Structuring the Space of Sociotechnical Alignment: A Specification Framework and Systematic Literature Review

Esra Dönmez and **Agnieszka Falenska**
Institute for Natural Language Processing, University of Stuttgart
Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart
esra.doenmez@ims.uni-stuttgart.de

Abstract

Sociotechnical alignment concerns the *social desirability of AI behavior* and is thus inherently ***normative***, not merely technical. While NLP research increasingly addresses its technical aspects, it often leaves underspecified what such "social desirability" entails. We argue that this reflects a fundamental gap: the absence of a systematic way to specify how sociotechnical alignment defines, justifies, and evaluates socially desirable AI behavior.

To address this gap, we introduce a human-centered framework for specifying sociotechnical alignment. We draw on social-scientific accounts of sociobehavioral desirability to ground the basis for behavioral desirability judgments and use this framework to analyze how alignment is specified in practice. Our systematic literature review identifies recurring patterns: normative concepts grounding desirability judgments are often unspecified or conflated with alignment targets for (desired) system behavior, target populations are underdefined, and design choices are rarely theoretically justified. These findings point to a lack of conceptual specificity that limits cumulative progress. We therefore offer recommendations that link social-scientific frameworks to alignment design choices, supporting more conceptually precise approaches to sociotechnical alignment.

1 Introduction

As AI systems become embedded in social contexts, they must meet not only functional goals, such as producing accurate recommendations, but also *normative expectations* about how they should *"behave"* in society, for example, by being fair and respectful in their interactions (Selbst et al., 2019; Weidinger et al., 2024). This broader challenge of specifying, modeling, and evaluating socially desirable AI behavior across social contexts is referred to as *sociotechnical alignment*.

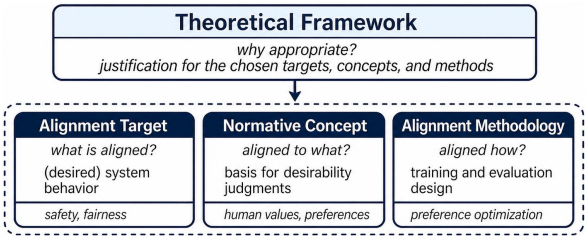

Figure 1: Framework for specifying sociotechnical alignment: overarching theoretical framework at the top, with three dimensions below: alignment target, normative concept, and alignment methodology, each with examples at the bottom.

Despite growing work on sociotechnical alignment, the literature still leaves "social desirability" underspecified. First, "alignment" is often used as a catch-all term for heterogeneous, socially contested objectives, ranging from user engagement to social acceptability (Gabriel and Ghazavi, 2023). Second, existing work often moves between different ways of specifying what AI systems should "align to", ranging from normative concepts (e.g., human values) to more instrumental proxies (e.g., user preferences or stakeholder requirements), without specifying the reasoning behind these choices (Ji et al., 2025b). This ambiguity is worsened when alignment is framed primarily in terms of "human values", blurring technical questions of system behavior with normative disputes over whose values matter and how they are interpreted (Gabriel, 2020; Sorensen et al., 2024). Work on pluralistic alignment further shows that evaluation criteria (how alignment is assessed) and relevant populations (whose judgments are considered) cannot be treated as universal (Rastogi et al., 2025). Taken together, these issues point to a deeper conceptual gap: **the absence of a systematic way to specify how sociotech-**

nical alignment defines, justifies, and evaluates socially desirable AI behavior.

The present work addresses this gap by structuring sociotechnical alignment along four interrelated dimensions (Figure 1): the *alignment target* specifying desired system behavior, the *normative concept* defining what the system aligns to, the *alignment methodology* for modeling and evaluation, and the *theoretical framework* justifying these choices across contexts and populations. We ground the normative-concept dimension in established social-scientific frameworks for conceptualizing sociobehavioral desirability (§3). We identify three key normative concepts—*values*, *moral judgments*, and *socially enforced norms*—and examine how they are formalized across levels of analysis, from individual-level preferences to universal normative claims. This synthesis provides a basis for applying these concepts in alignment settings and for relating them to evaluations of AI behavior across contexts and populations.

Next, we use the four dimensions to examine how sociotechnical alignment is specified in practice. To this end, we construct a corpus of 281 sociotechnical alignment papers[1] from the ACL Anthology[2] by filtering for publications with alignment- and normativity-related keywords, manually labeling titles and abstracts with alignment targets, and conducting targeted readings to identify, where possible, theoretical frameworks and design choices underlying the alignment methodology (§4.1). The analysis of this corpus shows that sociotechnical alignment research focuses primarily on technical aspects and safety constraints, with remaining work on bias, fairness, and personalization, and comparatively little on population-level value and moral compatibility (§4.2). Moreover, normative concepts are often left unspecified or conflated with alignment targets; target populations are frequently underdefined; and design choices are rarely theoretically justified (§4.3 and §4.4).

Taken together, our findings show why a four-dimensional sociotechnical alignment specification is needed: the problems we identify are not isolated omissions, but recurring disconnects between ***what is being aligned, for whom, by which methods, and on what normative grounds***. Building on this diagnosis, we derive practical recommendations for alignment research (§5). We argue for theory-grounded approaches, in which normative assumptions are explicit, consistently reflected across data, modeling, and evaluation, and appropriately scoped to represented populations. To support this, we provide a mapping of theoretical frameworks onto practical alignment decisions, linking settings such as personalization, moral conflict, and cross-cultural deployment to appropriate forms of normative grounding. These recommendations aim to support cumulative progress by making alignment approaches more comparable and reliably deployable.

[1]The corpus and the code used to build it are available at https://sociotechnicalalignment.github.io/.

[2]ACL Anthology allows for large-scale analysis within a manageable scope. Extending the analysis to other venues is a direction for future work.

2 Related Surveys

Broad alignment overviews. Wang et al. (2024f) situate alignment historically, trace its conceptual origins, examine its mathematical foundations, and review major alignment methods. Fernandes et al. (2023) survey the use of human feedback in natural language generation, and Jiang et al. (2025b) review preference learning for LLMs, covering feedback sources and formats, preference modeling, and evaluation. Together, these surveys capture the scope of alignment research, but they generally treat normativity at a high level. They show which alignment problems are being studied without much focus on the sociotechnical aspects.

Sociotechnical perspectives. These surveys frame alignment as an ongoing process of specification, interaction, and evaluation on technical and normative goals. Kirk et al. (2023) survey feedback learning for subjective preferences and values, reflecting on prior work and future directions. Terry et al. (2024) conceptualize interactive alignment across task specification, process, and evaluation, arguing that alignment should be understood across the full human–AI interaction loop rather than only through model optimization. Lu et al. (2025) review practical safety mechanisms and evaluation challenges for deployed LLMs, highlighting open problems in oversight, value pluralism, robustness, and continuous alignment. Ji et al. (2025b) propose a forward and backward view of alignment: systems are aligned forward through training and backward through governance, organized around four principles: Robustness, Interpretability,

Controllability, and Ethicality (RICE). Sharma et al. (2025b) focus on emerging issues such as multimodal alignment, cultural fairness, and low-latency optimization. However, while these works broaden alignment beyond purely technical optimization, they still leave a gap in explaining how research concretely specifies and justifies the social desirability of AI behavior.

Position of this survey. This survey complements prior overviews by focusing on how socially desirable AI behavior is specified in practice. Whereas existing surveys synthesize alignment methods, feedback mechanisms, training paradigms, and system-level challenges, we examine how sociotechnical alignment research specifies alignment and justifies its choices. We therefore treat sociotechnical alignment not only as a problem of optimization, but also as one of specifying, grounding, and evaluating social desirability in NLP.

3 Normative Concepts of Sociobehavioral Desirability

Social-scientific work generally treats *sociobehavioral desirability* through related but distinct concepts—*values*, *moral evaluations*, and *social norms*—whose meanings and effects vary by context. *Values* (e.g., equality) are relatively abstract, transsituational goals or priorities that orient preferences and evaluation across situations (Parsons, 1951; Schwartz, 1992; Schwartz et al., 2012; Ponizovskiy et al., 2019). *Norms* (e.g., disclosing sources) are context- and role-specific expectations about appropriate conduct, maintained through socialization, coordination, and sanctioning (Durkheim, 1982; Goffman, 1959; Bicchieri, 2006). *Morals* (e.g., lying is wrong) are judgments about right and wrong that carry stronger obligation and boundary-drawing force than many everyday norms, thereby legitimizing praise, blame, and social inclusion or exclusion (Durkheim, 1912; Haidt, 2001, 2012). Across these levels, values articulate broad *ends and priorities*, norms specify *appropriate means* of conduct in particular settings, and morals supply *evaluative standards* that render some expectations as right or wrong (Durkheim, 1912; Haidt, 2001).

Social-scientific accounts further situate these concepts across different levels of social organization. At the *individual level*, value hierarchies help explain preferences and action, while norms enter as expectations attached to roles and interaction settings (Schwartz, 2011, 2012; Witte et al., 2020; Durkheim, 1982; Goffman, 1959). Moral judgments distinguish ordinary norm violation from conduct experienced as obligatory, blameworthy, or unacceptable (Haidt, 2001; Graham et al., 2009, 2011). At the *cultural level*, the literature generally describes cultures less as possessing wholly different value structures but as differing in the relative emphasis placed on the same value dimensions, which in turn shapes which norms become salient and how strongly they are enforced (Schwartz, 2006, 1992; Stieger and Lewetz, 2016; Leite et al., 2021; Vignoles et al., 2018; De Giorgis and Gangemi, 2024; Durkheim, 1982; Goffman, 1959). At the *universal or normative level*, some traditions posit minimal moral baselines centered on harm avoidance, fairness, trust, and social cooperation, while still allowing substantial variation in how such concerns are interpreted and prioritized across groups (Gert, 2005; Gert et al., 2006; Haidt and Joseph, 2004; Graham et al., 2013; Schwartz, 1992, 2012). Here, *universality* concerns the assumption of broadly shared human in-

Level of analysis	Frameworks
Individual (micro)	Rokeach's Value Survey (RVS); Social Value Orientation (SVO); Kohlberg's Theory of Moral Development (TMD)
Individual ↔ Universal	Schwartz's Theory of Basic Human Values (TBHV); Moral Foundations Theory (MFT)
Cultural (macro)	Hofstede's Cultural Dimensions Theory (CDT)
Cultural ↔ Universal	Inglehart's Post-Materialist Thesis (PMT); World Values Survey (WVS) (incl. Inglehart–Welzel dimensions); Human Development Sequence (HDS)
Universal (normative)	Gert's Common Morality
Individual ↔ Cultural (bridging typology)	de Groot and Steg (2008) (egoistic–altruistic–biospheric value orientations); Sheldon and Nichols (2009) (ecocentrism vs. anthropocentrism); Bicchieri (2006)'s social norms framework; Norm Activation Model (NAM); Value–Belief–Norm (VBN)

Table 1: Selection of widely used frameworks for conceptualizing and/or operationalizing values, norms, and morals across levels of analysis (a more complete version in Appendix Table 4). Here, *level of analysis* refers to the main social or conceptual scale a framework addresses, from individual preferences and moral judgments to cultural patterns and universal normative claims. Bridging frameworks connect more than one level.

Keywords	Stemmed Keywords
alignment, value, moral, feedback, reinforcement, preferences, sociotechnical	align, value, moral, feedback, reinforc, prefer, sociotech

Table 2: Keywords used to retrieve articles from ACL. Highlighted keywords were left unstemmed, and stemming as well as the keyword *norm* were excluded due to many irrelevant matches.

tuitions, whereas *normativity* concerns how those intuitions are defined and made socially binding by groups, cultures, or individuals.

A selective set of frameworks illustrates how these distinctions are typically conceptualized in the literature (for a more comprehensive overview, see Table 1). For values, Schwartz's Theory of Basic Human Values is widely used to model recurring value priorities and trade-offs across individuals and cultures (Schwartz, 1992, 2012; Schwartz et al., 2012). For moral judgment, Moral Foundations Theory describes recurrent dimensions of moral concern and variation in how groups weight them (Haidt and Joseph, 2004; Haidt, 2001, 2007, 2012; Graham et al., 2013). For norms, Bicchieri (2006)'s social norms framework is especially relevant because it explains norms in terms of empirical expectations, normative expectations, and conditional preferences sustained by anticipated approval or disapproval. In this broader social-science view, these frameworks do not collapse into a single account of alignment; rather, they mark different levels at which sociobehavioral desirability is theorized and measured.

In summary, **sociobehavioral desirability is a layered and context-sensitive outcome of the interaction among values, norms, and moral judgments.** *Values orient preferences*, *norms define expected conduct in specific contexts*, and *moral judgments determine which expectations are treated as right or wrong*. Accordingly, assessments of acceptable behavior vary across individuals and cultures, even where they appeal to broader normative concerns such as harm avoidance, fairness, trust, and social cooperation.

4 Sociotechnical Alignment in NLP

In this section, we survey sociotechnical alignment literature in the ACL Anthology using the four-dimensional framework introduced above. We examine the *alignment target* specifying (desired) system behavior, the *normative concept* defining what the system aligns to, the *alignment methodology* for modeling and evaluation, and the *theoretical framework* justifying these choices. To this end, we first construct a dedicated corpus from the ACL Anthology and then analyze how sociobehavioral desirability is specified, operationalized, and justified in alignment research.

4.1 Building the Corpus

We extracted all papers from ACL Anthology[3] from 2022 to 2025[4] and then applied a multi-stage filtering procedure, which we describe below.

Automatic filtering We removed entries with no abstracts and filtered for abstracts in English. We then matched the remaining 37260 paper abstracts against our keyword list in Table 2, retaining papers with at least two matches as a candidate set of alignment-relevant papers, resulting in a total of 1089 papers (see Kirk et al. (2023) for a similar survey procedure)[5]. The keywords can be used in different senses, therefore, we performed an additional step of automatic filtering by creating topic clusters. This resulted in 20 clusters via topic modeling the abstracts using BERTopic with UMAP embeddings (detailed algorithms with exact parameters in Appendix §B.1). We then conservatively removed all the clusters that clearly focus on topics other than alignment, such as medical NLP and code generation, which resulted in 790 remaining papers (details in Appendix §B.1). Even this reduced set remained broad and contained many false positives.

Manual annotation To eliminate remaining false positives, we employed three annotators to read paper titles and abstracts (details in Appendix §B.2). The annotators read all 790 titles and abstracts and identified 497 papers with an alignment focus. They then annotated these papers based on abstracts and titles on dimensions of **(a)** sociotechnical focus, **(b)** modality, **(c)** language, and **(d)** alignment target. We retained papers with sociotechnical focus and text modality, yielding a final set of 281 papers for analysis in our survey.

[3] https://aclanthology.org/

[4] The general-purpose LLM era; cut-off: 05/08/2025.

[5] We matched on abstract to increase coverage. In a sample of 100 retrieved articles, we determined #keywords ≥ 2 best balanced relevancy with the size of the retrieved set.

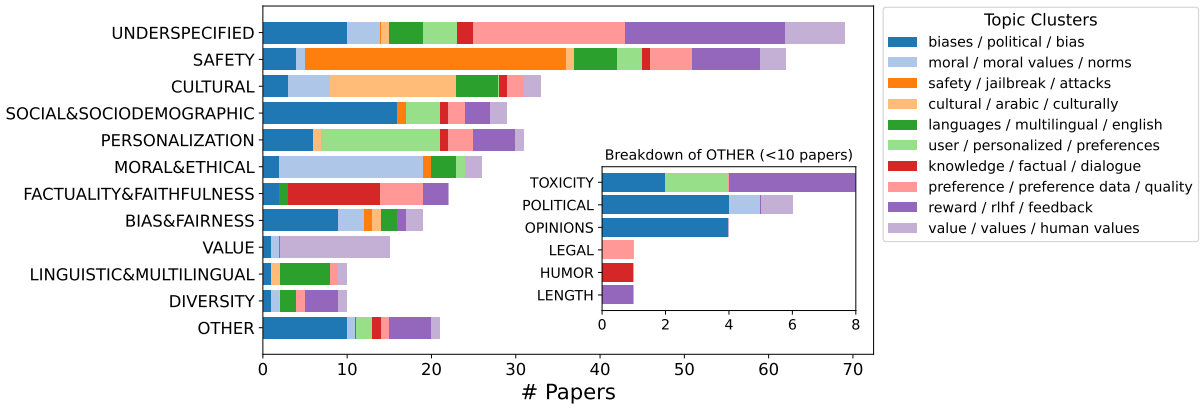

Figure 2: Distribution of annotated alignment targets across the 281-paper corpus, with automatically derived topic clusters overlaid within each manually-annotated target. Bar lengths show the number of papers per annotated target, and colored segments indicate the corresponding cluster composition from abstracts. The inset breaks down the heterogeneous OTHER category (<10 papers per subcategory).

Assigning Alignment Targets To capture each paper's sociotechnical alignment target, annotators assigned free-form labels based on the abstract, producing an initial set of 25 distinct labels (details in Appendix §B.2). We then merged semantically overlapping labels into broader categories to improve consistency; for example, *sexism*, *offensiveness*, *hate*, and *toxicity* were consolidated under *toxicity*. Papers could receive multiple labels, and the annotators were instructed to use *underspecified* when an abstract did not clearly state an alignment target.

Identifying Topical Clusters Annotation alone could not fully resolve target underspecification and terminological ambiguity. To better understand how alignment is framed in the literature, we supplemented manual annotation with topic modeling. We applied TF–IDF vectorization and Non-negative Matrix Factorization (NMF), using the scikit-learn implementations of both methods, to paper abstracts (details in Appendix §B.3).

Manual target labels were used to map alignment targets, while topic clusters identified broader target framings. We then conducted qualitative reading to examine how papers specified the framework's four dimensions: alignment target, normative concept, modeling or evaluation methodology, and theoretical framework. We closely read all papers in the value, moral, and cultural normative compatibility framings, where sociobehavioral desirability was most explicit. For the remaining framings, we targeted representative papers and papers exhibiting terminological ambiguity or target underspecification. We treated a design choice as theoretically justified when it was linked to a named theoretical framework, social-scientific construct, or grounded argument about target behavior or population appropriateness. This analysis supports Sections 4.2–4.4.

4.2 The Sociotechnical Alignment Landscape: Alignment Targets, Normative Concepts, and Operationalization

Sociotechnical alignment research in NLP spans diverse targets and problems. The manually assigned target labels (y-axis of Figure 2) reveal four main patterns. **Target underspecification** is common: *underspecified* is the most frequent label in the corpus. **Skewed target distribution** is also apparent: the literature focuses heavily on *safety alignment*, followed by mid-sized clusters such as *cultural alignment*, *sociodemographic factors*, and *bias and fairness*, with a long tail including *humor* and *verbosity*. **Values and morals are often treated as targets**: papers frequently model *values* and *morals* as alignment targets rather than as cross-cutting normative concepts. Finally, **terminological ambiguity** is widespread: annotators reported substantial variation in the use of key terms such as *values*, *safety*, and *fairness*.

Further probing these patterns, our complementary topic modeling yielded ten distinct clusters shown in the legend of Figure 2 and in Table 3. The overlaps and divergences between these clusters and the manually assigned target labels in Figure 2 indicate that the two approaches are com-

plementary. Most automatically assigned cluster labels broadly align with the corresponding manually labeled targets. Where they diverge, the mismatches often point to deeper ***conceptual underspecification***, ***terminological ambiguity***—for example, omitted key terms or shared keywords used with different meanings across papers—or differences in what each approach captures, such as the method, the alignment target, or the normative concept emphasized.

To better communicate how sociotechnical alignment is operationalized in the literature, we consolidate these automatically identified clusters under broader framing categories (cf. Table 3): Alignment as ***(a)*** *preference aggregation and optimization* (T1, T2), ***(b)*** *bias and fairness assessment* (T0), ***(c)*** *safety constraints* (T3), ***(d)*** *personalization* (T5), ***(e)*** *multilingual and cross-lingual transfer* (T7), ***(f)*** *epistemic reliability* (T8), and ***(g)*** *value and moral compatibility* (T4, T6, T9). We summarize these target framings with attention to the *normative concepts* invoked, the *methodologies* used to model and evaluate them, and the *theoretical frameworks* used to justify these choices.

(A) Alignment as Preference Aggregation and Optimization

A central alignment framing (75 papers) treats alignment as **optimizing behavior under (often multiple and conflicting) objectives** (e.g., helpfulness, harmlessness, honesty) (Guo et al., 2024b; Zhou et al., 2024b,a). The works in this category are generally domain-agnostic and are shaped by the availability of *preference* data, which is used as the normative concept through which social alignment is modeled. They treat *"social preferences"* as heterogeneous and group-dependent signals, where the technical goal (optimization) is driven by human label variation in evaluative judgments (Kumar et al., 2025; Zeng et al., 2024). However, the mechanisms that guide these social preferences (values, morals, and norms) are typically modeled indirectly (via preference ratings) rather than through explicit constructs or theories (Kumar et al., 2025; Nath et al., 2025b).

(B) Alignment as Public-Facing Bias and Fairness Assessment

A prominent bias-focused cluster (45 papers) frames alignment as **the measurement and mitigation of biases in LMs** (e.g., demographic and political), with emphasis on how prompts elicit differential behavior across groups and how models reproduce stereotypes (e.g., racial and gendered) in generated text (Huang and Xiong, 2024; Liu et al., 2024a). Rather than treating alignment solely as reward maximization, this line of work operationalizes alignment through public-facing fairness criteria—quantifying bias, comparing outcomes across demographic groups, and auditing politically charged or sensitive settings where harms manifest in model outputs (Nghiem et al., 2024; Potter et al., 2024). These works frame sociotechnical alignment in terms of *disparities in model outputs across demographic, cultural, or political groups* under controlled prompting and evaluation settings, connecting those differences to potential equity and representational harms (Asad et al., 2025; Sun et al., 2025). Although the overall alignment target is often clear, the normative basis for these socially contested objectives frequently remains implicit.

Cluster: Alignment Framing (Topic Clusters)	#
0: Bias, veracity, and stereotyping effects	45
1: RLHF, reward modeling, and learning from human feedback	39
2: Preference data quality and optimization, and response ranking	36
3: Jailbreaks, red-teaming, and safety against harmful behavior	32
4: Value theory operationalization and value datasets/principles	28
5: Personalization, user modeling, and persona conditioning	25
6: Moral belief evaluation and moral psychology benchmarks	22
7: Multilingual instruction tuning and cross-lingual alignment	21
8: Factuality, consistency, and knowledge calibration in dialogue	17
9: Cultural alignment and cross-cultural evaluation	16

Table 3: Distribution of thematic clusters in the corpus ($N = 281$). Cluster labels were derived from top TF–IDF terms in paper abstracts, and capture how the literature frames sociotechnical alignment (cf. Appendix Table 9).

(C) Alignment as Safety Constraints

Another dominant framing (32 papers) defines alignment in terms of **safety constraints**, often under adversarial prompting and distribution shift (Xu et al., 2024e; Zhao et al., 2024). This includes *inference-time defenses* and *training-time safety specialization* (Liu et al., 2024b; Zhou et al., 2025a). These works often treat sociotechnical alignment as a *system property* (robust refusal/safe completion) that must hold even when users behave strategically (Cao et al., 2024a; Guo et al., 2025b). Here, *safety* broadly covers issues ranging from universal safety governance to socially contested topics such as bias and ethics. The *"social"* component typically appears as *normative safety*

standards and *policy constraints*, operationalized through harmful-prompt sets, refusal targets, and safety scoring rubrics (Xu et al., 2024a; Ji et al., 2025a). However, these normative safety standards are often inherited from existing datasets without systematic evaluation of their validity or representativeness (e.g., whether human ethical judgments can be translated into AI behavioral standards, or whether safety criteria developed for one population remain appropriate in another population or deployment context).

(D) Alignment as Personalization and User-Centered Evaluation

A large theme (25 papers) frames alignment as **fit to specific users, intents, or contexts**: *user-centered benchmarks* that define "aligned" responses relative to user intents and multi-intent task structure, and *persona/personalization mechanisms* that condition generation on user history or persona representations to improve subjective fit (Wang et al., 2024d; Wu et al., 2025b). These works treat alignment as *interactional and situated*: model behavior should be coherent with user profiles and goals (Baskar et al., 2025; Balepur et al., 2025). The social-science connection is often implicit (user modeling, preference heterogeneity) rather than anchored in explicit theory (Guan et al., 2025; Castricato et al., 2025).

(E) Multilingual and Cross-Lingual Alignment Transfer

Several papers (21 papers) investigate **alignment across languages**, frequently treating language as a proxy for community difference and measurement shift (Kravchenko et al., 2025; Agarwal et al., 2024). Works often discuss cross-lingual instruction tuning, transfer of safety/morality constructs, and failure modes driven by resource imbalance or annotation mismatch (Chen et al., 2025a; Shen et al., 2024). Social science engagement here varies: some work explicitly interprets language-community differences through cultural or normative lenses, while other work treats multilinguality primarily as a technical generalization problem (Jinnai, 2024; Dang et al., 2024).

(F) Alignment as Epistemic Reliability in Dialogue: Factuality and Consistency

A distinct knowledge-centered cluster (17 papers) frames alignment as **epistemic reliability in interaction**: producing responses that remain factually grounded, internally consistent, and stable across multi-turn dialogue (Xue et al., 2023; Bi et al., 2025). The normative concept in this framing is epistemic reliability, based on the implicit assumption that users want system properties such as factuality, coherence, and consistency, though this is rarely framed as social alignment. Work in this cluster focuses on reducing conversational contradictions, improving faithful expression and retrieval of factual knowledge, and aligning a model's internal state with its responses (Liang et al., 2024; Wu et al., 2024a). Evaluation is dialogue-centric, assessing whether responses remain factually correct across turns and cohere with earlier context and the model's stated beliefs, rather than relying only on one-shot truthfulness scores (Xue et al., 2023; Xie et al., 2025b).

(G) Alignment as Value, Moral, and Cultural Normative Compatibility

The works discussed below frame sociotechnical alignment in terms of values, morality, and cultural norms, along with variation in values and morality across cultures, and aim to model these factors to define and evaluate desirable model behavior. In this literature, these normative concepts are sometimes modeled as alignment to the value, moral, or normative profiles of particular populations, and sometimes as properties of the model itself, expressed as a relatively stable value profile.

Value alignment. Some papers (28 papers; cf. T4 in Table 3) study "values" through constructs associated with moral philosophy and social psychology to examine value adherence, tradeoffs, or priorities (Yao et al., 2024a; Ye et al., 2025; Liu et al., 2025d). Values are typically operationalized through taxonomies, survey-style items, or inferred value profiles, and evaluated by comparing model outputs to human or theory-derived value distributions. Other work uses "values" more loosely, for example through high-dimensional representations or benchmarked national value profiles (Cahyawijaya et al., 2025; Ju et al., 2025).

Moral alignment. Other work (22 papers; cf. T6 in Table 3) studies moral alignment through moral theories, questionnaires, and benchmark tasks, sometimes drawing on cognitive-developmental accounts and moral-psychology constructs to probe models' "moral beliefs", disagreement, and cross-lingual or cross-group variation (Liu et al., 2024f; Abdulhai et al.,

2024; Davani et al., 2024; Alvarez Nogales and Araque, 2024). Morality is commonly operationalized through dilemma judgments, moral-foundation labels, or acceptability ratings, and evaluated against human judgments. But some papers operationalize morality more loosely through moral dialogue, self-correction, or policy consistency (Sun et al., 2023; Rao et al., 2023).

Cross-cultural compatibility. Cross-cultural work (16 papers; cf. T9 in Table 3) sometimes draws on cultural-dimensions theory to study cultural differences and compare model behavior with human cultural distributions (Masoud et al., 2025; Wang et al., 2024g; Cao et al., 2024c). Culture is operationalized through cultural dimension surveys, cultural scenarios, or comparisons to population-level response distributions. Other work operationalizes culture through localization, cultural sensitivity, or adaptation pipelines (Banerjee et al., 2025; Huang et al., 2024a).

In summary, sociotechnical alignment research is concentrated on broad technical optimization and safety constraints, while much of the remaining work addresses behavioral target-focused problems such as bias, fairness, personalization, consistency, and factuality.

4.3 Terminological Imprecision

Across the literature, papers rarely specify the dimensions of sociotechnical alignment with precision. Terms such as "values," "social alignment," "safety," or "fairness" often signal an alignment concern without clarifying what, concretely, is being aligned. In particular, papers often leave unclear: (i) the alignment target specifying system behavior; (ii) the normative concept invoked; (iii) how the concept is translated into modeling or evaluation practices for that behavioral target; (iv) which population is treated as relevant and why; and (v) how far claims generalize beyond the observed data. This imprecision is most visible in underspecified contributions and in work on safety, and value, moral and cultural compatibility.

4.4 Conceptual and Theoretical Gaps

This imprecision becomes most consequential when papers explicitly invoke values, morals, or cultural norms as alignment targets. We therefore examine how these concepts are defined, operationalized, and theoretically justified.

Normative concept is explicitly grounded in well-known theoretical frameworks. For *values*, grounding appears in work that treats values as measurable behavioral dispositions and evaluates them using instruments adapted from the social sciences, such as surveys, inventories, and tests (Zhou et al., 2023; Qiu et al., 2022; Long et al., 2025; Jiang et al., 2025a; Liu et al., 2025d; Choi et al., 2025; Yao et al., 2025). For *morals*, grounding appears in work that treats morality as a measurable pattern of judgment and evaluates it using dilemma-based benchmarks, moral-psychology frameworks, standardized probing instruments, and disagreement-sensitive measures (Wu and Deng, 2025; Marraffini et al., 2024; Guo et al., 2024a; Greco et al., 2025; Davani et al., 2024; Falk and Lapesa, 2025; Abdulhai et al., 2024; Fraser et al., 2022; Ramezani and Xu, 2023; Liu et al., 2024f; Alvarez Nogales and Araque, 2024). For *culture*, grounding appears in work that treats cultural norms as measurable patterns of appropriateness and variation and evaluates them using cultural-dimension frameworks, survey-conditioned dialogue settings, and culture-targeted adaptation signals (Xu et al., 2025c; Arora et al., 2023; Wang et al., 2024g; Masoud et al., 2025; Cao et al., 2024c; Liu et al., 2025a).

Normative concept is made explicit but definition is context- and method-specific. Here, works point to a theoretical lineage or framework while keeping the normative concept more method- or context-specific. This includes organizational "value" compliance framed as alignment to explicitly stated organizational values, modeling "values" as patterns in a model's internal representations, nation-grounded benchmarking framed around "values", psychometrics-*motivated* value measurement, generative approaches that examine how language influences value expression, culture-specific ethical instruction datasets, and moral-knowledge injection framed as an inspired operationalization of "moral values" (Mittal et al., 2025; Cahyawijaya et al., 2025; Ju et al., 2025; Han et al., 2025; Ye et al., 2025; Hu et al., 2024a; Hakim et al., 2025; Zhang et al., 2024c).

Normative concept is loosely defined. For *values*, this underspecification appears in work where "values" functions as an umbrella term for heterogeneous, and at times mismatched, objectives defined primarily by what can be induced, probed,

or optimized. These include natural-language value statements, implicit preference signals from human feedback, constraints defined as socially beneficial, representational directions in activation space, population-level or demographic correlates, causal value graphs, prioritized rule sets, stakeholder priorities, and context-specific safety or morality dimensions (Bang et al., 2023; Liu et al., 2022; Ammanabrolu et al., 2022; Jin et al., 2025; Dong et al., 2025; Lee et al., 2024c; Liu et al., 2024c; Al Ali and Libovický, 2024; Kang et al., 2025; Lu et al., 2024b; Xu et al., 2025a; Marco et al., 2025; Shetty et al., 2025; Huang et al., 2024b). For *morals*, underspecification appears in work that operationalizes morality as universal right-or-wrong judgment, norm compliance, policy consistency, feedback-driven self-correction, values or lessons extracted from arguments or stories, association-based comparisons of moral values, or aggregate collective reasoning, without grounding these targets in an explicit theoretical framework (Leteno et al., 2025; Guan et al., 2022; Emelin et al., 2021; Sun et al., 2023; Liu et al., 2025b; Rao et al., 2023; Paulissen and Wendt, 2023; Hobson et al., 2024; Xiang et al., 2025a; Yuan et al., 2025). For *culture*, underspecification appears in work that operationalizes cultural norms through sensitivity, community-grounded competence, or localization pipelines, rather than treating them as an explicitly theorized target (Banerjee et al., 2025; Feng et al., 2025; Magdy et al., 2025; Nacar et al., 2025; Huang et al., 2024a; Hosseinbeigi et al., 2025; Hu et al., 2022; Acquaye et al., 2024; Keleg, 2025; Pandey et al., 2025).

5 Implications and Practical Recommendations

Sociotechnical alignment research in NLP is constrained less by a lack of methods than by a lack of conceptual precision within and across studies. Throughout the literature, alignment targets, normative concepts, methodologies, and population-specific considerations are often left implicit or treated interchangeably, producing mismatches across works. Clearer alignment research should therefore specify what is being aligned, to what, and for whom, and should ideally ground these choices in explicit justificatory frameworks. We distill these observations into practical recommendations for future alignment research.

Separate the behavioral target from the normative concepts. Values, morals, norms, and preferences should not be treated as interchangeable labels and not be confused with socially desirable behavior. Likewise, safety, toxicity, fairness, misinformation, or personalization are alignment targets, not normative concepts. Work on sociotechnical alignment should state both the behavioral target being modeled and the normative concept invoked. If the normative concept is defined only in terms of general preferences, work should avoid using stronger normative terms such as *human values* unless they justify that translation.

Ground concepts in theory. When papers invoke values, morals, or (cultural) norms, they should anchor these concepts in an explicit theoretical measurement tradition. If preferences, annotations, or benchmark labels are used as proxies, that translation and its limits should be made clear.

Make the intended layer of alignment explicit across data, modeling, and evaluation. Sociobehavioral desirability operates across multiple levels, including individual preferences, situational norms, cultural patterns, and broader normative baselines. Alignment research should specify which level it targets and ensure that data collection, optimization, and evaluation match that level. This includes stating whose judgments are represented, how disagreement is handled, and what scope of generalization is being claimed.

We next provide illustrative guidelines for selecting theoretical frameworks to support this process. These examples are not exhaustive; the appropriate framework depends on modeling and deployment contexts across behavioral targets and populations.

Personalization and interaction design (user-level alignment). At the micro/individual level, *Rokeach's Value Survey (RVS)* distinguishes *terminal* values (desired end-states) from *instrumental* values (preferred means), helping systems separate what users want from how they prefer to achieve it. In design, this can support targeted interactions, though it risks oversimplifying preferences or over-personalizing when values are inferred too confidently or treated as stable. *Social Value Orientation (SVO)* helps model tradeoffs between self- and other-regarding outcomes, relevant for allocation, negotiation, and recommendation settings. *Kohlberg's Theory of Moral De-*

velopment can guide explanations and refusals by varying justification style, such as rule-, harm-, or rights-based rationales, without changing the underlying policy.

Value conflict detection and robust framing (moralized topics). Frameworks bridging individual and more general moral structure are especially useful for handling controversial or morally loaded requests. *Schwartz's Theory of Basic Human Values (TBHV)* provides a cross-context vocabulary for common tensions (openness to change vs. conservatism; self-transcendence vs. self-enhancement) and can be used to (a) surface tradeoffs explicitly and (b) parameterize interaction defaults (e.g., cautious vs. exploratory, privacy-first vs. convenience-first) in a way that is legible and empirically grounded. *Moral Foundations Theory (MFT)* is well-suited to auditing the assistant's framing and tone: it helps identify when responses systematically privilege certain moral rhetorics (e.g., care/fairness) while alienating others (e.g., authority/loyalty/sanctity/liberty), which can drive perceived misalignment even when factual content is correct.

Cross-cultural deployment and localization (population-level alignment). For globally deployed systems, macro-level cultural frameworks can help anticipate differences in expectations and acceptability. *Hofstede's Cultural Dimensions Theory* can guide localization of interaction norms, such as directness, authority, and uncertainty avoidance, and identify where one-size-fits-all alignment may fail. *Inglehart's Post-Materialist Thesis* explains macro-level variation in priorities, such as security and stability versus autonomy and self-expression, motivating stratified evaluation designs. Because these frameworks are coarse and population-level, they should be treated as hypotheses rather than deterministic rules for individual users.

Bridging individual and collective goals (impact-sensitive domains). The *De Groot & Steg* egoistic–altruistic–biospheric typology can be practical for systems that advise on consumption, travel, energy, or sustainability-related decisions: it can structure how recommendations trade off cost/time (egoistic), social welfare (altruistic), and environmental impact (biospheric), and can be used for preference elicitation without requiring deep political or ideological inference.

Universal constraints and governance (normative floor). At the universal/normative level, *Gert's Common Morality* can define a "moral floor" of constraints, such as avoiding serious harm, deception, and coercion, that remain stable across personalization and localization. This supports layered alignment: a universal normative floor, locale-specific interaction norms, and user-level personalization within those bounds.

Practical integration in alignment research. Across these stages, the frameworks are most actionable when used to: (a) design elicitation instruments (e.g., surveys and preference annotation) and stakeholder models (e.g., end users, impacted non-users, and communities), (b) translate "values" into interaction policies and constraints, (c) construct stratified evaluations that test for alignment variation across value profiles and cultures, and (d) support tradeoff documentation.

6 Conclusion

This work introduced a four-dimensional framework for specifying sociotechnical alignment and used it to analyze how alignment is specified in NLP research. Focusing on English ACL Anthology work enabled systematic comparison, with multilingual, cross-disciplinary, and non-ACL extensions left to future work. Our analysis revealed recurring underspecification across alignment targets, normative concepts, populations, and the scope of generalization, as well as a broader limitation: alignment research often lacks a shared basis for comparison, making claims difficult to justify and compare across studies.

The identified lack of specification has important consequences. When normative assumptions remain implicit, socially contingent judgments may be presented as universal, particularly in systems deployed across diverse contexts. This motivates the need for future work to explicitly articulate what is aligned, to what, for whom, and under which assumptions. At the same time, greater conceptual precision does not resolve which normative concepts should guide system behavior: social-scientific frameworks can clarify how values, norms, and moral judgments are defined and operationalized, but not which of them should govern system behavior. In this sense, rather than prescribing a single account of alignment, the proposed framework is intended to make these choices more transparent and comparable.

Limitations

Scope limitations of keyword- and cluster-based retrieval. Our literature identification strategy relies on keyword queries and cluster-based filtering. While this approach supports systematic coverage at scale, it may under-represent relevant work that uses alternative terminology. We adopted this setup because it enabled high recall and, when combined with automatic and manual filtering, reduced both false negatives and false positives. As a result, some contributions, particularly those situated at disciplinary boundaries, may still have been omitted.

English-language restriction. We limited inclusion to English-language publications. This supported more consistent screening and synthesis, but it may also have introduced linguistic and regional bias by excluding studies published in other languages, including work shaped by local regulatory contexts, cultural norms, and region-specific system deployments. However, because the venues in our sample primarily publish in English, we believe this criterion remains broadly representative of the field.

Community and venue selection effects. Our review primarily reflects the emphases of the ACL community most visible in the sociotechnical alignment discourse as operationalized by our search strategy. Adjacent communities may conceptualize alignment differently, prioritize distinct alignment targets, or employ alternative methodological standards. Consequently, the themes and gaps we identify should be interpreted as contingent on this community sampling rather than exhaustive of all relevant works.

Ethical Considerations

This paper examines how sociotechnical alignment research conceptualizes and operationalizes sociobehavioral desirability. Because concepts such as values, moral judgments, and social norms shape decisions about what AI systems should model, optimize, or evaluate, their use carries ethical and political implications. Treating these concepts as interchangeable or self-evident can obscure whose perspectives are being prioritized and which assumptions are embedded in alignment design.

A central ethical motivation of this work is to make these assumptions more visible. Underspecified alignment targets, normative bases, and evaluation criteria can make socially contingent judgments appear universal, especially when relevant populations are left implicit. This is particularly consequential for systems deployed across heterogeneous social contexts, where values, norms, and judgments of appropriate behavior may differ across groups. Our framework is intended to support greater transparency about these choices, rather than to prescribe a single correct account of sociobehavioral desirability.

We also emphasize that conceptual precision does not eliminate normative disagreement. Social-scientific frameworks can help clarify what kind of normative concept is being invoked, how it is operationalized, and for whom it is meant to apply, but they cannot determine by themselves which values, norms, or moral judgments should govern system behavior. Such decisions require explicit justification and, where appropriate, participatory and accountable forms of deliberation.

Finally, our recommendations should not be read as encouraging the uncritical modeling of any population-level preference, norm, or value. Social norms and group preferences may themselves encode exclusion, domination, or harm. For this reason, alignment design should distinguish between describing existing sociobehavioral patterns and endorsing them as appropriate targets for system behavior.

Acknowledgments

We acknowledge the support of the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK, Ministry of Science, Research and the Arts Baden-Württemberg under Az. 33-7533-9 19/54/5) in Künstliche Intelligenz & Gesellschaft: Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) and the support by the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empiri-*

cal Methods in Natural Language Processing, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics.

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, Miami, Florida, USA. Association for Computational Linguistics.

Christabel Acquaye, Haozhe An, and Rachel Rudinger. 2024. Susu box or piggy bank: Assessing cultural commonsense knowledge between Ghana and the US. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9483–9502, Miami, Florida, USA. Association for Computational Linguistics.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

Ibrahim Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are generative language models multicultural? a study on Hausa culture and emotions using ChatGPT. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 98–106, Bangkok, Thailand. Association for Computational Linguistics.

Adnan Al Ali and Jindřich Libovický. 2024. How gender interacts with political values: A case study on Czech BERT models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8200–8210, Torino, Italia. ELRA and ICCL.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Yazeed Alnumay, Alexandre Barbet, Anna Bialas, William Darling, Shaan Desai, Joan Devassy, Kyle Duffy, Stephanie Howe, Olivia Lasche, Justin Lee, Anirudh Shrinivason, and Jennifer Tracey. 2025. Command R7B Arabic: a small, enterprise-focused, multilingual, and culturally aware Arabic LLM. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 126–135, Vienna, Austria. Association for Computational Linguistics.

Anny D. Alvarez Nogales and Oscar Araque. 2024. Moral disagreement over serious matters: Discovering the knowledge hidden in the perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 67–77, Torino, Italia. ELRA and ICCL.

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9954–9972, Bangkok, Thailand. Association for Computational Linguistics.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Narjis Asad, Nihar Ranjan Sahoo, Rudra Murthy, Swaprava Nath, and Pushpak Bhattacharyya. 2025. "you are beautiful, body image stereotypes are ugly!" BIStereo: A benchmark to

measure body image stereotypes in language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24471–24496, Vienna, Austria. Association for Computational Linguistics.

Abinew Ali Ayele, Esubalew Alemneh Jalew, Adem Chanie Ali, Seid Muhie Yimam, and Chris Biemann. 2024. Exploring boundaries and intensities in offensive and hate speech: Unveiling the complex spectrum of social media discourse. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 167–178, Torino, Italia. ELRA and ICCL.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico. Association for Computational Linguistics.

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Whose boat does it float? improving personalization in preference tuning via inferred user personas. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3371–3393, Vienna, Austria. Association for Computational Linguistics.

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025. Navigating the cultural kaleidoscope: A hitchhiker's guide to sensitivity in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7580–7617, Albuquerque, New Mexico. Association for Computational Linguistics.

Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab, and Pascale Fung. 2023. Enabling classifiers to make judgements explicitly aligned with human values. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 311–325, Toronto, Canada. Association for Computational Linguistics.

Logan Barnhart, Reza Akbarian Bafghi, Stephen Becker, and Maziar Raissi. 2025. Aligning to what? limits to RLHF based alignment. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7556–7591, Albuquerque, New Mexico. Association for Computational Linguistics.

Sarvesh Baskar, Manas Gaur, Srinivasan Parthasarathy, and Tanmay Tulsidas Verlekar. 2025. (CPER) from guessing to asking: An approach to resolving persona knowledge gap in LLMs during multi-turn conversations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 435–447, Albuquerque, USA. Association for Computational Linguistics.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. 2025. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10280–10300, Vienna, Austria. Association for Computational Linguistics.

Claudia Biancotti, Carolina Camassa, Andrea Coletta, Oliver Giudice, and Aldo Glielmo. 2025. Chat bankman-fried: an exploration of LLM alignment in finance. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 1–22, Abu Dhabi, UAE. Association for Computational Linguistics.

Cristina Bicchieri. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, Cambridge.

Yuyan Bu, Liangyu Huo, Yi Jing, and Qing Yang. 2025. Beyond excess and deficiency: Adaptive length bias mitigation in reward models for RLHF. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3091–3098, Albuquerque, New Mexico. Association for Computational Linguistics.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2025. High-dimension human value representation in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5330, Albuquerque, New Mexico. Association for Computational Linguistics.

Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. 2024a. Defending against alignment-breaking attacks via robustly aligned LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10542–10560, Bangkok, Thailand. Association for Computational Linguistics.

Meng Cao, Lei Shu, Lei Yu, Yun Zhu, Nevan Wichers, Yinxiao Liu, and Lei Meng. 2024b. Enhancing reinforcement learning with dense rewards from language model critic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9119–9138, Miami, Florida, USA. Association for Computational Linguistics.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024c. Bridging cultural nuances in dialogue agents through cultural value surveys. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945, St. Julian's, Malta. Association for Computational Linguistics.

Yuanpu Cao, Bochuan Cao, and Jinghui Chen. 2024d. Stealthy and persistent unalignment on large language models via backdoor injections. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4920–4935, Mexico City, Mexico. Association for Computational Linguistics.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. PERSONA: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368, Abu Dhabi, UAE. Association for Computational Linguistics.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

Khaoula Chehbouni, Megha Roshan, Emmanuel Ma, Futian Wei, Afaf Taik, Jackie Cheung, and Golnoosh Farnadi. 2024. From representational harms to quality-of-service harms: A case study on llama 2 safety safeguards. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15694–15710, Bangkok, Thailand. Association for Computational Linguistics.

Hongyu Chen and Seraphina Goldfarb-Tarrant. 2025. Safer or luckier? LLMs as safety evaluators are not robust to artifacts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19750–19766, Vienna, Austria. Association for Computational Linguistics.

Kuang-Ming Chen, Jenq-Neng Hwang, and Hung-yi Lee. 2025a. InstructionCP: A simple yet effective approach for transferring large language models to target languages. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 1–6, Vienna, Austria. Association for Computational Linguistics.

Quan Ze Chen, Kevin Feng, Chan Young Park, and Amy X Zhang. 2025b. SPICA: Retrieving scenarios for pluralistic in-context alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 748–765, Vienna, Austria. Association for Computational Linguistics.

Xiusi Chen, Hongzhi Wen, Sreyashi Nag, Chen Luo, Qingyu Yin, Ruirui Li, Zheng Li, and

Wei Wang. 2024a. IterAlign: Iterative constitutional alignment of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1423–1433, Mexico City, Mexico. Association for Computational Linguistics.

Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 2024b. Are U a joke master? pun generation via multi-stage curriculum learning towards a humor LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 878–890, Bangkok, Thailand. Association for Computational Linguistics.

Yanjun Chen, Dawei Zhu, Yirong Sun, Xinghao Chen, Wei Zhang, and Xiaoyu Shen. 2024c. The accuracy paradox in RLHF: When better reward models don't yield better language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2980–2989, Miami, Florida, USA. Association for Computational Linguistics.

Zhi-Yuan Chen, Shiqi Shen, Guangyao Shen, Gong Zhi, Xu Chen, and Yankai Lin. 2024d. Towards tool use alignment of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1382–1400, Miami, Florida, USA. Association for Computational Linguistics.

Pengyu Cheng, Yifan Yang, Jian Li, Yong Dai, Tianhao Hu, Peixin Cao, Nan Du, and Xiaolong Li. 2024. Adversarial preference optimization: Enhancing your alignment via RM-LLM game. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3705–3716, Bangkok, Thailand. Association for Computational Linguistics.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2024. InstructEval: Towards holistic evaluation of instruction-tuned large language models. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 35–64, St. Julian's, Malta. Association for Computational Linguistics.

Sooyung Choi, Jaehyeok Lee, Xiaoyuan Yi, Jing Yao, Xing Xie, and JinYeong Bak. 2025. Unintended harms of value-aligned LLMs: Psychological and empirical insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31742–31768, Vienna, Austria. Association for Computational Linguistics.

Nicholas E. Corrado, Julian Katz-Samuels, Adithya M Devraj, Hyokun Yun, Chao Zhang, Yi Xu, Yi Pan, Bing Yin, and Trishul Chilimbi. 2025. AutoMixAlign: Adaptive data mixing for multi-task preference optimization in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20234–20258, Vienna, Austria. Association for Computational Linguistics.

Ander Corral, Ixak Sarasua Antero, and Xabier Saralegi. 2025. Pipeline analysis for developing instruct LLMs in low-resource languages: A case study on Basque. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12636–12655, Albuquerque, New Mexico. Association for Computational Linguistics.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13134–13156, Miami, Florida, USA. Association for Computational Linguistics.

Souvik Das and Rohini Srihari. 2024. UNIWIZ: A unified large language model orchestrated wizard for safe knowledge grounded conversations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1749–1762, Bangkok, Thailand. Association for Computational Linguistics.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Stefano De Giorgis and Aldo Gangemi. 2024. That's all folks: A kg of values as commonsense social norms and behaviors. In *Value Engineering in Artificial Intelligence*, pages 11–27, Cham. Springer Nature Switzerland.

Weilong Dong, Xinwei Wu, Renren Jin, Shaoyang Xu, and Deyi Xiong. 2025. CONTRANS: Weak-to-strong alignment engineering via concept transplantation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4130–4148, Abu Dhabi, UAE. Association for Computational Linguistics.

Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. SteerLM: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11275–11288, Singapore. Association for Computational Linguistics.

Esra Dönmez and Agnieszka Falenska. 2025. "I understand your perspective": LLM persuasion through the lens of communicative action theory. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15312–15327, Vienna, Austria. Association for Computational Linguistics.

Pengfan Du, Sirui Liang, Baoli Zhang, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. ZhuJiu-knowledge: A fairer platform for evaluating multiple knowledge types in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 194–206, Mexico City, Mexico. Association for Computational Linguistics.

Jinhao Duan, Xinyu Zhao, Zhuoxuan Zhang, Eunhye Grace Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, Chenxi Yuan, Qi Long, Ying Ding, Tianlong Chen, and Kaidi Xu. 2025. GuideLLM: Exploring LLM-guided conversation with applications in autobiography interviewing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5558–5588, Albuquerque, New Mexico. Association for Computational Linguistics.

Shitong Duan, Xiaoyuan Yi, Peng Zhang, Yan Liu, Zheng Liu, Tun Lu, Xing Xie, and Ning Gu. 2024. Negating negatives: Alignment with human negative samples via distributional dispreference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1012–1042, Miami, Florida, USA. Association for Computational Linguistics.

Émile Durkheim. 1912. *The Elementary Forms of Religious Life*. Free Press, New York. English translation commonly cited; original French edition 1912.

Émile Durkheim. 1982. *The Rules of Sociological Method*. The Free Press, New York.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2022. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2056–2077, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2025. Mining the uncertainty patterns of humans and models in the annotation of moral foundations and human values. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22898–22921, Vienna, Austria. Association for Computational Linguistics.

Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, and Pengfei Liu. 2024. Reformatted alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 574–597, Miami, Florida, USA. Association for Computational Linguistics.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. CulFiT: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22413–22430, Vienna, Austria. Association for Computational Linguistics.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:1643–1668.

Kathleen C. Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? probing delphi's moral philosophy. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 26–42, Seattle, U.S.A. Association for Computational Linguistics.

Tingchen Fu, Yupeng Hou, Julian McAuley, and Rui Yan. 2025. Unlocking decoding-time controllability: Gradient-free multi-objective alignment with contrastive prompts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 366–384, Albuquerque, New Mexico. Association for Computational Linguistics.

Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. On the relationship between truth and political bias in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9018, Miami, Florida, USA. Association for Computational Linguistics.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437.

Iason Gabriel and Vafa Ghazavi. 2023. The challenge of value alignment: From fairer algorithms to ai safety. In *Oxford Handbook of Digital Ethics*. Oxford University Press.

Michael Galarnyk, Agam Shah, Dipanwita Guhathakurta, Poojitha Nandigam, and Sudheer Chava. 2025. How inclusively do LMs perceive social and moral norms? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4859–4869, Albuquerque, New Mexico. Association for Computational Linguistics.

Ben Ganon, Alon Zolfi, Omer Hofman, Inderjeet Singh, Hisashi Kojima, Yuval Elovici, and Asaf Shabtai. 2025. DIESEL: A lightweight inference-time safety enhancement for language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23870–23890, Vienna, Austria. Association for Computational Linguistics.

Bernard Gert. 2005. *Morality: Its Nature and Justification*. Oxford University Press.

Bernard Gert, Charles M. Culver, and K. Danner Clouser. 2006. *Bioethics: A systematic approach*. Oxford University Press.

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY.

Reshmi Gopalakrishna Pillai, Antske Fokkens, and Wouter van Atteveldt. 2025. Engagement-driven persona prompting for rewriting news

tweets. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8612–8622, Abu Dhabi, UAE. Association for Computational Linguistics.

J. Graham, Brian A. Nosek, J. Haidt, Ravi Iyer, Spassena P. Koleva, and Peter H. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Candida Maria Greco, Lucio La Cava, Lorenzo Zangari, and Andrea Tagarelli. 2025. Exploring LLMs' ability to spontaneously and conditionally modify moral expressions through text manipulation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18047–18070, Vienna, Austria. Association for Computational Linguistics.

Judith I. M. de Groot and Linda Steg. 2008. Value orientations to explain beliefs related to environmental significant behavior: How to measure egoistic, altruistic, and biospheric value orientations. *Environment and Behavior*, 40(3):330–354.

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5069–5087, Seattle, United States. Association for Computational Linguistics.

Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. 2025. A survey on personalized Alignment—The missing piece for large language models in real-world applications. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5313–5333, Vienna, Austria. Association for Computational Linguistics.

Honglin Guo, Kai Lv, Qipeng Guo, Tianyi Liang, Zhiheng Xi, Demin Song, Qiuyinzhe Zhang, Yu Sun, Kai Chen, Xipeng Qiu, and Tao Gui. 2025a. CritiQ: Mining data quality criteria from human preferences. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16240–16261, Vienna, Austria. Association for Computational Linguistics.

Rongchen Guo, Isar Nejadgholi, Hillary Dawkins, Kathleen C. Fraser, and Svetlana Kiritchenko. 2024a. Adaptable moral stances of large language models on sexist content: Implications for society and gender discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19548–19564, Miami, Florida, USA. Association for Computational Linguistics.

Weiyang Guo, Jing Li, Wenya Wang, Yu Li, Daojing He, Jun Yu, and Min Zhang. 2025b. MTSA: Multi-turn safety alignment for LLMs through multi-round red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26424–26442, Vienna, Austria. Association for Computational Linguistics.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024b. Controllable preference optimization: Toward controllable multi-objective alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1437–1454, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan Haidt. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834.

Jonathan Haidt. 2007. The new synthesis in moral psychology. *Science*, 316(5827):998–1002.

Jonathan Haidt. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage Books.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Sulthan Abiyyu Hakim, Rizal Setya Perdana, and Tirana Noor Fatyanosa. 2025. Anak baik: A low-cost approach to curate Indonesian ethical and unethical instructions. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 52–62, Online. Association for Computational Linguistics.

Sil Hamilton. 2024. Detecting mode collapse in language models via narration. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 65–72, St. Julian's, Malta. Association for Computational Linguistics.

Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. 2025. Value portrait: Assessing language models' values through psychometrically and ecologically valid items. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17119–17159, Vienna, Austria. Association for Computational Linguistics.

Yupu Hao, Pengfei Cao, Zhuoran Jin, Huanxuan Liao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Evaluating personalized tool-augmented LLMs from the perspectives of personalization and proactivity. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21897–21935, Vienna, Austria. Association for Computational Linguistics.

Þórir Hrafn Harðarson, Hrafn Loftsson, and Stefán Ólafsson. 2025. Aligning language models for Icelandic legal text summarization. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 241–251, Tallinn, Estonia. University of Tartu Library.

Rima Hazra, Sayan Layek, Somnath Banerjee, and Soujanya Poria. 2024. Safety arithmetic: A framework for test-time safety alignment of language models by steering parameters and activations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21759–21776, Miami, Florida, USA. Association for Computational Linguistics.

Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. 2024. Whose emotions and moral sentiments do language models reflect? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6611–6631, Bangkok, Thailand. Association for Computational Linguistics.

David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032, Miami, Florida, USA. Association for Computational Linguistics.

Jixiang Hong, Quan Tu, Changyu Chen, Gao Xing, Ji Zhang, and Rui Yan. 2024. CycleAlign: Iterative distillation from black-box LLM to white-box models for better human alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14596–14609, Bangkok, Thailand. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, MohammadAli SeifKashani, Javad Seraj, Fatemeh Taherinezhad, Ali Nafisi, Fatemeh Nadi, Iman Barati, Hosein Hasani, Mostafa Amiri, and Mostafa Masoudi. 2025. Matina: A culturally-aligned Persian language model using multiple LoRA experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20874–20889, Vienna, Austria. Association for Computational Linguistics.

Alexander Hoyle, Lorena Calvo-Bartolomé, Jordan Boyd-Graber, and Philip Resnik. 2025. ProxAnn: Use-oriented evaluations of topic

models and document clustering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15872–15897, Vienna, Austria. Association for Computational Linguistics.

I-Hung Hsu, Zifeng Wang, Long Le, Lesly Miculicich, Nanyun Peng, Chen-Yu Lee, and Tomas Pfister. 2024. CaLM: Contrasting large and small language models to verify grounded generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12782–12803, Bangkok, Thailand. Association for Computational Linguistics.

Brian Hu, Bill Ray, Alice Leung, Amy Summerville, David Joy, Christopher Funk, and Arslan Basharat. 2024a. Language models are alignable decision-makers: Dataset and application to the medical triage domain. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 213–227, Mexico City, Mexico. Association for Computational Linguistics.

Chi Hu, Yimin Hu, Hang Cao, Tong Xiao, and JingBo Zhu. 2024b. Teaching language models to self-improve by learning from language feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6090–6101, Bangkok, Thailand. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, and Fei Liu. 2023. DecipherPref: Analyzing influential factors in human preference judgments via GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8344–8357, Singapore. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024a. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

James Y. Huang, Sailik Sengupta, Daniele Bonadiman, Yi-An Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. 2025. DeAL: Decoding-time alignment for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26280–26300, Vienna, Austria. Association for Computational Linguistics.

Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, and Dahua Lin. 2024b. Flames: Benchmarking value alignment of LLMs in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4551–4591, Mexico City, Mexico. Association for Computational Linguistics.

Shijia Huang, Jianqiao Zhao, Yanyang Li, and Liwei Wang. 2023. Learning preference model for LLMs via automatic preference data generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9187–9199, Singapore. Association for Computational Linguistics.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng,

Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. 2025a. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31983–32016, Vienna, Austria. Association for Computational Linguistics.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Hua Xu, Aidan O'Gara, Kwan Ng, Brian Tse, Jie Fu, Stephen Mcaleer, Yanfeng Wang, Mingchuan Yang, Yunhuai Liu, Yizhou Wang, Song-Chun Zhu, Yike Guo, Yaodong Yang, and Wen Gao. 2025b. Ai alignment: A contemporary survey. *ACM Comput. Surv.*, 58(5).

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025c. Enhancing persona consistency for LLMs' role-playing using persona-aware contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26221–26238, Vienna, Austria. Association for Computational Linguistics.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025a. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria. Association for Computational Linguistics.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2025b. A survey on human preference learning for aligning large language models. *ACM Comput. Surv.*, 58(6).

Weipeng Jiang, Zhenting Wang, Juan Zhai, Shiqing Ma, Zhengyu Zhao, and Chao Shen. 2025c. An optimizable suffix is worth a thousand templates: Efficient black-box jailbreaking without affirmative phrases via LLM as optimizer. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5471–5483, Albuquerque, New Mexico. Association for Computational Linguistics.

Haoran Jin, Meng Li, Xiting Wang, Zhihao Xu, Minlie Huang, Yantao Jia, and Defu Lian. 2025. Internal value alignment in large language models through controlled value vector activation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27347–27371, Vienna, Austria. Association for Computational Linguistics.

Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 48–64, Bangkok, Thailand. Association for Computational Linguistics.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168, Mexico City, Mexico. Association for Computational Linguistics.

Chengyi Ju, Weijie Shi, Chengzhong Liu, Jiaming Ji, Jipeng Zhang, Ruiyuan Zhang, Jiajie Xu, Yaodong Yang, Sirui Han, and Yike Guo. 2025. Benchmarking multi-national value alignment for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20042–20058, Vienna, Austria. Association for Computational Linguistics.

Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, and Zilong Zheng. 2025. Are the values of LLMs structurally aligned with humans? a causal perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23147–23161, Vienna, Austria. Association for Computational Linguistics.

Karine Karine and Benjamin Marlin. 2025. Using LLMs to improve RL policies in personalized health adaptive interventions. In *Proceedings of the Second Workshop on Patient-Oriented*

Language Processing (CL4Health), pages 137–147, Albuquerque, New Mexico. Association for Computational Linguistics.

Amr Keleg. 2025. LLM alignment for the Arabs: A homogenous culture or diverse ones. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 1–9, Albuquerque, New Mexico. Association for Computational Linguistics.

Aisha Khatun and Daniel Brown. 2023. Reliability check: An analysis of GPT-3's response to sensitive topics and prompt wording. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 73–95, Toronto, Canada. Association for Computational Linguistics.

Ahrii Kim. 2025. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165, Vienna, Austria. Association for Computational Linguistics.

Sean Kim and Hyuhng Joon Kim. 2025. A dual-layered evaluation of geopolitical and cultural bias in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 580–595, Vienna, Austria. Association for Computational Linguistics.

Hannah Rose Kirk, Andrew M. Bean, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore. Association for Computational Linguistics.

Frederic Kirstein, Terry Ruas, and Bela Gipp. 2025. Is my meeting summary good? estimating quality with a multi-LLM evaluator. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 561–574, Abu Dhabi, UAE. Association for Computational Linguistics.

Sungho Ko, Hyunjin Cho, Hyungjoo Chae, Jinyoung Yeo, and Dongha Lee. 2024. Evidence-focused fact summarization for knowledge-augmented zero-shot question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10636–10651, Miami, Florida, USA. Association for Computational Linguistics.

George Kour, Itay Nakash, Michal Shmueli-Scheuer, and Ateret Anaby Tavor. 2025. Think again! the effect of test-time compute on preferences, opinions, and beliefs of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 639–660, Vienna, Austria. Association for Computational Linguistics.

Andrian Kravchenko, Yurii Paniv, and Nazarii Drushchak. 2025. UAlign: LLM alignment benchmark for the Ukrainian language. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 36–44, Vienna, Austria (online). Association for Computational Linguistics.

Anna Kruspe. 2024. Musical ethnocentrism in large language models. In *Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA)*, pages 62–68, Oakland, USA. Association for Computational Linguistics.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A. Smith, and Hannaneh Hajishirzi. 2025. ComPO: Community preferences for language model personalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8246–8279, Albuquerque, New Mexico. Association for Computational Linguistics.

Tharindu Kumarage, Ninareh Mehrabi, Anil Ramakrishna, Xinyan Zhao, Richard Zemel, Kai-Wei Chang, Aram Galstyan, Rahul Gupta, and Charith Peris. 2025. Towards safety reasoning in LLMs: AI-agentic deliberation for policy-embedded CoT data creation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22694–22715, Vienna, Austria. Association for Computational Linguistics.

Oh Joon Kwon, Daiki E. Matsunaga, and Kee-Eung Kim. 2024. GDPO: Learning to directly align language models with diversity using GFlowNets. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17120–17139, Miami, Florida, USA. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Gihun Lee, Minchan Jeong, Yujin Kim, Hojung Jung, Jaehoon Oh, SangMook Kim, and Se-Young Yun. 2024a. BAPO: Base-anchored preference optimization for overcoming forgetting in large language models personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6804–6820, Miami, Florida, USA. Association for Computational Linguistics.

Huije Lee, Hoyun Song, Jisu Shin, Sukmin Cho, SeungYoon Han, and Jong C. Park. 2024b. Towards effective counter-responses: Aligning human preferences with strategies to combat online trolling. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11670–11686, Miami, Florida, USA. Association for Computational Linguistics.

Jaewook Lee, Junseo Jang, Oh-Woog Kwon, and Harksoo Kim. 2025. Small changes, big impact: How manipulating a few neurons can drastically alter LLM aggression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23478–23505, Vienna, Austria. Association for Computational Linguistics.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024c. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand. Association for Computational Linguistics.

Suhyun Lee and ChangHeon Han. 2025. Sentimatic: Sentiment-guided automatic generation of preference datasets for customer support dialogue system. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 120–128, Albuquerque, USA. Association for Computational Linguistics.

Ângela Leite, Ana Ramires, Diogo Guedes Vidal, Hélder Fernando Pedrosa e Sousa, Maria Alzira Pimenta Dinis, and Alexandra Fidalgo. 2021. Hierarchical cluster analysis of human value priorities and associations with subjective well-being, subjective general health, social life, and depression across europe. *Social Sciences*.

Maria Lerner, Florian Dorner, Elliott Ash, and Naman Goel. 2024. Whose preferences? differences in fairness preferences and their impact on the fairness of AI utilizing human feedback. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9403–9425, Bangkok, Thailand. Association for Computational Linguistics.

Thibaud Leteno, Irina Proskurina, Antoine Gourru, Julien Velcin, Charlotte Laclau, Guillaume Metzler, and Christophe Gravier. 2025. HISTOIRESMORALES: A French dataset for assessing moral alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2590–2612, Albuquerque, New Mexico. Association for Computational Linguistics.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. 2025a. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 11214–11232, Vienna, Austria. Association for Computational Linguistics.

Shusheng Li, Jiale Li, Yifei Qu, Xinwei Shi, Yanliang Guo, Ziyi He, Yubo Wang, and Wenjun Tan. 2025b. Semantic-eval : A semantic comprehension evaluation framework for large language models generation without training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9675–9690, Vienna, Austria. Association for Computational Linguistics.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025c. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria. Association for Computational Linguistics.

Yanhao Li, Hongshen Chen, Heng Zhang, Zhiwei Ge, Tianhao Li, Sulong Xu, and Guibo Luo. 2025d. Unraveling the mystery: Defending against jailbreak attacks via unearthing real intention. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8374–8384, Abu Dhabi, UAE. Association for Computational Linguistics.

Yu Li, Devamanyu Hazarika, Di Jin, Julia Hirschberg, and Yang Liu. 2024. From pixels to personas: Investigating and modeling self-anthropomorphism in human-robot dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9695–9713, Miami, Florida, USA. Association for Computational Linguistics.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in LLMs for hallucination mitigation. In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pages 44–58, Bangkok, Thailand. Association for Computational Linguistics.

Baihan Lin, Djallel Bouneffouf, Guillermo Cecchi, and Kush Varshney. 2024. Towards healthy AI: Large language models need therapists too. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 61–70, Mexico City, Mexico. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024a. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025a. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Guangliang Liu, Zhiyu Xue, Xitong Zhang, Rongrong Wang, and Kristen Johnson. 2025b. Smaller large language models can do moral self-correction. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 56–65, Albuquerque, New Mexico. Association for Computational Linguistics.

Qianying Liu, Katrina Qiyao Wang, Fei Cheng, and Sadao Kurohashi. 2025c. 7 points to Tsinghua but 10 points to ? assessing large language models in agentic multilingual national bias. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26430–26442, Vienna, Austria. Association for Computational Linguistics.

Quan Liu, Zhenhong Zhou, Longzhu He, Yi Liu, Wei Zhang, and Sen Su. 2024b. Alignment-enhanced decoding: Defending jailbreaks via token-level adaptive refining of probability distributions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2802–2816, Miami, Florida, USA. Association for Computational Linguistics.

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics:*

NAACL 2022, pages 241–252, Seattle, United States. Association for Computational Linguistics.

Siyang Liu, Trisha Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. 2024c. The generation gap: Exploring age bias in the value systems of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19617–19634, Miami, Florida, USA. Association for Computational Linguistics.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024d. Aligning large language models with human preferences through representation engineering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10619–10638, Bangkok, Thailand. Association for Computational Linguistics.

Xin Liu, Farima Fatahi Bayat, and Lu Wang. 2024e. Enhancing language model factuality via activation-based confidence calibration and guided decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10436–10448, Miami, Florida, USA. Association for Computational Linguistics.

Xuelin Liu, Pengyuan Liu, and Dong Yu. 2025d. What's the most important value? INVP: INvestigating the value priorities of LLMs through decision-making in social scenarios. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4725–4752, Abu Dhabi, UAE. Association for Computational Linguistics.

Xuelin Liu, Yanfei Zhu, Shucheng Zhu, Pengyuan Liu, Ying Liu, and Dong Yu. 2024f. Evaluating moral beliefs across LLMs through a pluralistic framework. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4740–4760, Miami, Florida, USA. Association for Computational Linguistics.

Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, and James Kwok. 2025e. Mixture of insighTful experts (MoTE): The synergy of reasoning chains and expert mixtures in self-alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3022–3038, Vienna, Austria. Association for Computational Linguistics.

Saüc Abadal Lloret, Shehzaad Dhuliawala, Keerthiram Murugesan, and Mrinmaya Sachan. 2024. Towards aligning language models with textual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20240–20266, Miami, Florida, USA. Association for Computational Linguistics.

Do Xuan Long, Kenji Kawaguchi, Min-Yen Kan, and Nancy Chen. 2025. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2526–2547, Abu Dhabi, UAE. Association for Computational Linguistics.

Haoran Lu, Luyang Fang, Ruidong Zhang, Xinliang Li, Jiazhang Cai, Huimin Cheng, Lin Tang, Ziyu Liu, Zeliang Sun, Tao Wang, Yingchuan Zhang, Arif Hassan Zidan, Jinwen Xu, Jincheng Yu, Meizhi Yu, Hanqi Jiang, Xilin Gong, Weidi Luo, Bolun Sun, Yongkai Chen, Terry Ma, Shushan Wu, Yifan Zhou, Junhao Chen, Haotian Xiang, Jing Zhang, Afrar Jahin, Wei Ruan, Ke Deng, Yi Pan, Peilong Wang, Jiahui Li, Zhengliang Liu, Lu Zhang, Lin Zhao, Wei Liu, Dajiang Zhu, Xin Xing, Fei Dou, Wei Zhang, Chao Huang, Rongjie Liu, Mengrui Zhang, Yiwen Liu, Xiaoxiao Sun, Qin Lu, Zhen Xiang, Wenxuan Zhong, Tianming Liu, and Ping Ma. 2025. Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges.

Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2023. Towards boosting the open-domain chatbot with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4060–4078, Toronto, Canada. Association for Computational Linguistics.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. 2024a. Eliminat-

ing biased length reliance of direct preference optimization via down-sampled KL divergence. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1047–1067, Miami, Florida, USA. Association for Computational Linguistics.

Xinyu Lu, Bowen Yu, Yaojie Lu, Hongyu Lin, Haiyang Yu, Le Sun, Xianpei Han, and Yongbin Li. 2024b. SoFA: Shielded on-the-fly alignment via priority rule following. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7108–7136, Bangkok, Thailand. Association for Computational Linguistics.

Yi Luo, Zhenghao Lin, YuHao Zhang, Jiashuo Sun, Chen Lin, Chengjin Xu, Xiangdong Su, Yelong Shen, Jian Guo, and Yeyun Gong. 2024. Ensuring safe and high-quality outputs: A guideline library approach for language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1152–1197, Mexico City, Mexico. Association for Computational Linguistics.

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024. The potential and challenges of evaluating attitudes, opinions, and values in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12320–12341, Albuquerque, New Mexico. Association for Computational Linguistics.

Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand. Association for Computational Linguistics.

Guillermo Marco, Julio Gonzalo, and Víctor Fresno. 2025. The reader is the metric: How textual features and reader profiles explain conflicting evaluations of AI creative writing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25432–25449, Vienna, Austria. Association for Computational Linguistics.

Giovanni Franco Gabriel Marraffini, Andrés Cotton, Noe Fabian Hsueh, Axel Fridman, Juan Wisznia, and Luciano Del Corro. 2024. The greatest good benchmark: Measuring LLMs' alignment with utilitarian moral dilemmas. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21950–21959, Miami, Florida, USA. Association for Computational Linguistics.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.

Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. 2025. Rejected dialects: Biases against African American language in reward models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7468–7487, Albuquerque, New Mexico. Association for Computational Linguistics.

Kshitij Mishra, Manisha Burja, and Asif Ekbal. 2024. ABLE: Personalized disability support with politeness and empathy integration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22445–22470, Miami, Florida, USA. Association for Computational Linguistics.

Avni Mittal, Sree Hari Nagaralu, and Sandipan Dandapat. 2025. PROTECT: Policy-related organizational value taxonomy for ethical compliance and trust. In *Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025)*, pages 73–75, Vienna, Austria. Association for Computational Linguistics.

Samraj Moorjani, Adit Krishnan, Hari Sundaram, Ewa Maslowska, and Aravind Sankar. 2022. Audience-centric natural language generation via style infusion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1919–1932, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11241–11258, Albuquerque, New Mexico. Association for Computational Linguistics.

Omer Nacar, Serry Taiseer Sibaee, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S. Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, Mohamed Abdelkader, and Anis Koubaa. 2025. Towards inclusive Arabic LLMs: A culturally aligned benchmark in Arabic large language model evaluation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 387–401, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tahira Naseem, Guangxuan Xu, Sarathkrishna Swaminathan, Asaf Yehudai, Subhajit Chaudhury, Radu Florian, Ramón Astudillo, and Asim Munawar. 2024. A grounded preference model for LLM alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 151–162, Bangkok, Thailand. Association for Computational Linguistics.

Abhijnan Nath, Carine Graff, Andrei Bachinin, and Nikhil Krishnaswamy. 2025a. Frictional agent alignment framework: Slow down and don't break things. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11042–11089, Vienna, Austria. Association for Computational Linguistics.

Abhijnan Nath, Andrey Volozin, Saumajit Saha, Albert Aristotle Nanda, Galina Grunin, Rahul Bhotika, and Nikhil Krishnaswamy. 2025b. DPL: Diverse preference learning without a reference model. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3727–3747, Albuquerque, New Mexico. Association for Computational Linguistics.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé Iii. 2024. "you gotta be a doctor, lin" : An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7268–7287, Miami, Florida, USA. Association for Computational Linguistics.

Dang Nguyen, Jiuhai Chen, and Tianyi Zhou. 2024a. Multi-objective linguistic control of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4336–4347, Bangkok, Thailand. Association for Computational Linguistics.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024b. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–

304, Bangkok, Thailand. Association for Computational Linguistics.

Minhyeon Oh, Seungjoon Lee, and Jungseul Ok. 2025. Comparison-based active preference learning for multi-dimensional personalization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33145–33166, Vienna, Austria. Association for Computational Linguistics.

Inkit Padhi, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Manish Nagireddy, Pierre Dognin, and Kush R. Varshney. 2024. Value alignment from unstructured text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1083–1095, Miami, Florida, US. Association for Computational Linguistics.

Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. CULTURALLY YOURS: A reading assistant for cross-cultural content. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 208–216, Abu Dhabi, UAE. Association for Computational Linguistics.

Deepak Pandita, Tharindu Cyril Weerasooriya, Sujan Dutta, Sarah K. Luger, Tharindu Ranasinghe, Ashiqur R. KhudaBukhsh, Marcos Zampieri, and Christopher M. Homan. 2024. Rater cohesion and quality from a vicarious perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5149–5162, Miami, Florida, USA. Association for Computational Linguistics.

Bumjin Park, Jinsil Lee, and Jaesik Choi. 2025. Deontological keyword bias: The impact of modal expressions on normative judgments of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7277–7296, Vienna, Austria. Association for Computational Linguistics.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017, Bangkok, Thailand. Association for Computational Linguistics.

Talcott Parsons. 1951. *The Social System*. Free Press, Glencoe, IL.

Spencer Paulissen and Caroline Wendt. 2023. Lauri ingman at SemEval-2023 task 4: A chain classifier for identifying human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 193–198, Toronto, Canada. Association for Computational Linguistics.

Vladimir Ponizovskiy, Lusine Grigoryan, Ulrich Kühnen, and Klaus Boehnke. 2019. Social construction of the value–behavior relation. *Frontiers in Psychology*, Volume 10 - 2019.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden persuaders: LLMs' political leaning and their influence on voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.

Alexander Pugachev, Alena Fenogenova, Vladislav Mikhailov, and Ekaterina Artemova. 2025. REPA: Russian error types annotation for evaluating text generation and judgment capabilities. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 136–150, Vienna, Austria. Association for Computational Linguistics.

Weicong Qin, Yi Xu, Weijie Yu, Chenglei Shen, Ming He, Jianping Fan, Xiao Zhang, and Jun Xu. 2025. MAPS: Motivation-aware personalized search via LLM-driven consultation alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3039–3051, Vienna, Austria. Association for Computational Linguistics.

Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. 2022. Towards socially intelligent agents with mental state transition and human value. In *Proceedings of the 23rd Annual Meeting of the*

Special Interest Group on Discourse and Dialogue, pages 146–158, Edinburgh, UK. Association for Computational Linguistics.

Deng Qiyuan, Xuefeng Bai, Kehai Chen, Yaowei Wang, Liqiang Nie, and Min Zhang. 2025. Efficient safety alignment of large language models via preference re-ranking and representation-based reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31156–31171, Vienna, Austria. Association for Computational Linguistics.

Ravi Shanker Raju, Swayambhoo Jain, Bo Li, Jonathan Lingjie Li, and Urmish Thakker. 2024. Constructing domain-specific evaluation sets for LLM-as-a-judge. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 167–181, Miami, Florida, USA. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, Roma Patel, Ding Wang, Mark Diaz, Alicia Parrish, Aida Mostafazadeh Davani, Zoe Ashwood, Michela Paganini, Vinodkumar Prabhakaran, Verena Rieser, and Lora Aroyo. 2025. Whose view of safety? a deep DIVE dataset for pluralistic alignment of text-to-image models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Traian Rebedea, Leon Derczynski, Shaona Ghosh, Makesh Narsimhan Sreedhar, Faeze Brahman, Liwei Jiang, Bo Li, Yulia Tsvetkov, Christopher Parisien, and Yejin Choi. 2025. Guardrails and security for LLMs: Safe, secure and controllable steering of LLM applications. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 13–15, Vienna, Austria. Association for Computational Linguistics.

Qibing Ren, Chang Gao, Jing Shao, Junchi Yan, Xin Tan, Wai Lam, and Lizhuang Ma. 2024. CodeAttack: Revealing safety generalization challenges of large language models via code completion. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11437–11452, Bangkok, Thailand. Association for Computational Linguistics.

Virgile Rennard, Christos Xypolopoulos, and Michalis Vazirgiannis. 2025. Bias in the mirror : Are LLMs opinions robust to their own adversarial attacks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2128–2143, Vienna, Austria. Association for Computational Linguistics.

Ali Riahi Samani, Tianhao Wang, Kangshuo Li, and Feng Chen. 2025. Large language models with reinforcement learning from human feedback approach for enhancing explainable sexism detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6230–6243, Abu Dhabi, UAE. Association for Computational Linguistics.

Jesse Roberts, Kyle Moore, and Douglas Fisher. 2025. Do large language models learn human-like strategic preferences? In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pages 97–108, Vienna, Austria. Association for Computational Linguistics.

Corby Rosset, Guoqing Zheng, Victor Dibia, Ahmed Awadallah, and Paul Bennett. 2023. Axiomatic preference modeling for longform question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11445–11475, Singapore. Association for Computational Linguistics.

Amit Rozner, Barak Battash, Lior Wolf, and Ofir Lindenbaum. 2024. Knowledge editing in language models via adapted direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4761–4774, Miami, Florida, USA. Association for Computational Linguistics.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140, Bangkok, Thailand. Association for Computational Linguistics.

Shalom Schwartz. 2006. A theory of cultural value orientations: Explication and applications. *Comparative Sociology*, 5(2-3):137 – 182.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Academic Press, San Diego, CA.

Shalom H. Schwartz. 2011. Values: cultural and individual. In Fons J. R. van de Vijver, Athanasios Chasiotis, and Seger M.Editors Breugelmans, editors, *Fundamental Questions in Cross-Cultural Psychology*, page 463–493. Cambridge University Press.

Shalom H. Schwartz. 2012. An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2:11.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gümüş, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA. Association for Computing Machinery.

Nikhil Sharma, Kenton Murray, and Ziang Xiao. 2025a. Faux polyglot: A study on information disparity in multilingual large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8090–8107, Albuquerque, New Mexico. Association for Computational Linguistics.

Raghav Sharma, Manan Mehta, and Sai Tiger Raina. 2025b. Rlhf: A comprehensive survey for cultural, multimodal and low latency alignment methods.

Kennon Sheldon and Charles Nichols. 2009. Comparing democrats and republicans on intrinsic and extrinsic values. *Journal of Applied Social Psychology*, 39:589 – 623.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.

Anudeex Shetty, Amin Beheshti, Mark Dras, and Usman Naseem. 2025. VITAL: A new dataset for benchmarking pluralistic alignment in healthcare. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22954–22974, Vienna, Austria. Association for Computational Linguistics.

Taiwei Shi, Kai Chen, and Jieyu Zhao. 2024. Safer-instruct: Aligning language models with automated preference data. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7636–7651, Mexico City, Mexico. Association for Computational Linguistics.

Erjon Skenderi, Salla-Maaria Laaksonen, and Jukka Huhtamäki. 2025. Team conversational

AI: Introducing effervesce. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 502–513, Albuquerque, USA. Association for Computational Linguistics.

Feifan Song, Shaohang Wei, Wen Luo, Yuxuan Fan, Tianyu Liu, Guoyin Wang, and Houfeng Wang. 2025a. Well begun is half done: Low-resource preference alignment by weak-to-strong decoding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12654–12670, Vienna, Austria. Association for Computational Linguistics.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2025b. Multilingual blending: Large language model safety alignment evaluation with language mixture. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3433–3449, Albuquerque, New Mexico. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. 2024. Do LLMs plan like human writers? comparing journalist coverage of press releases with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21814–21828, Miami, Florida, USA. Association for Computational Linguistics.

Stefan Stieger and David Lewetz. 2016. Parent-child proximity and personality: basic human values and moving distance. *BMC Psychology*, 4.

Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2023. MoralDial: A framework to train and evaluate moral dialogue systems via moral discussions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2213–2230, Toronto, Canada. Association for Computational Linguistics.

Lihao Sun, Chengzhi Mao, Valentin Hofmann, and Xuechunzi Bai. 2025. Aligned but blind: Alignment increases implicit bias by reducing awareness of race. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22167–22184, Vienna, Austria. Association for Computational Linguistics.

Junya Takayama, Masaya Ohagi, Tomoya Mizumoto, and Katsumasa Yoshikawa. 2025. Persona-consistent dialogue generation via pseudo preference tuning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5507–5514, Abu Dhabi, UAE. Association for Computational Linguistics.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust LLMs: Aligning confidence with response quality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5984–5996, Bangkok, Thailand. Association for Computational Linguistics.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267, Miami, Florida, USA. Association for Computational Linguistics.

Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2024. Interactive ai alignment: Specification, process, and evaluation alignment.

William Thorne, Ambrose Robinson, Bohua Peng, Chenghua Lin, and Diana Maynard. 2024. Increasing the difficulty of automatically generated questions via reinforcement learning with synthetic preference for cost-effective cultural heritage dataset generation. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 450–462, Miami, USA. Association for Computational Linguistics.

Vivian L. Vignoles, Peter B. Smith, Maja Becker, and Matthew J. Easterbrook. 2018. In search of a pan-european culture: European values, beliefs, and models of selfhood in global perspective. *Journal of Cross-Cultural Psychology*, 49:868 – 887.

Anvesh Rao Vijjini, Rakesh R Menon, Jiayi Fu, Shashank Srivastava, and Snigdha Chaturvedi. 2024. SocialGaze: Improving the integration of human social norms in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16487–16506, Miami, Florida, USA. Association for Computational Linguistics.

Sahil Wadhwa, Chengtian Xu, Haoming Chen, Aakash Mahalingam, Akankshya Kar, and Divya Chaudhary. 2025. Northeastern uni at multilingual counterspeech generation: Enhancing counter speech generation with LLM alignment through direct preference optimization. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 19–28, Abu Dhabi, UAE. Association for Computational Linguistics.

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2024a. Learning personalized alignment for evaluating open-ended text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13274–13292, Miami, Florida, USA. Association for Computational Linguistics.

Haiyang Wang, Zhiliang Tian, Xin Song, Yue Zhang, Yuchen Pan, Hongkui Tu, Minlie Huang, and Bin Zhou. 2024b. Intent-aware and hate-mitigating counterspeech generation via dual-discriminator guided LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9131–9142, Torino, Italia. ELRA and ICCL.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024c. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655, Bangkok, Thailand. Association for Computational Linguistics.

Huanqian Wang, Yang Yue, Rui Lu, Jingxin Shi, Andrew Zhao, Shenzhi Wang, Shiji Song, and Gao Huang. 2025a. Model surgery: Modulating LLM's behavior via simple parameter editing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6337–6357, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024d. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA. Association for Computational Linguistics.

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. 2024e. RLHFPoison: Reward poisoning attack for reinforcement learning with human feedback in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2551–2570, Bangkok, Thailand. Association for Computational Linguistics.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025b. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21082–21107, Vienna, Austria. Association for Computational Linguistics.

Xinpeng Wang, Shitong Duan, Xiaoyuan Yi, Jing Yao, Shanlin Zhou, Zhihua Wei, Peng Zhang, Dongkuan Xu, Maosong Sun, and Xing Xie. 2024f. On the essence and prospect: an investigation of alignment approaches for big models. In *Proceedings of the Thirty-Third International*

Joint Conference on Artificial Intelligence, IJCAI '24.

Yuanfu Wang, Pengyu Wang, Chenyang Xi, Bo Tang, Junyi Zhu, Wenqiang Wei, Chen Chen, Chao Yang, Jingfeng Zhang, Chaochao Lu, Yijun Niu, Keming Mao, Zhiyu Li, Feiyu Xiong, Jie Hu, and Mingchuan Yang. 2025c. Adversarial preference learning for robust LLM alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21865–21881, Vienna, Austria. Association for Computational Linguistics.

Yue Wang, Haoke Zhang, Juntao Li, Jinxiong Chang, and Min Zhang. 2025d. $\mathcal{A}^3$: Automatic alignment framework for attributed text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28975–28990, Vienna, Austria. Association for Computational Linguistics.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024g. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

Yumeng Wang, Zhiyuan Fan, Qingyun Wang, Yi R. Fung, and Heng Ji. 2025e. CALM: Unleashing the cross-lingual self-aligning ability of language model question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2809–2817, Albuquerque, New Mexico. Association for Computational Linguistics.

Zekun Moore Wang, Shenzhi Wang, King Zhu, Jiaheng Liu, Ke Xu, Jie Fu, Wangchunshu Zhou, and Wenhao Huang. 2025f. PopAlign: Diversifying contrasting patterns for a more comprehensive alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28893–28921, Vienna, Austria. Association for Computational Linguistics.

Zihan Wang, XinZhang Liu, Shixuan Liu, Yitong Yao, Yunyao Huang, Mengxiang Li, Zhongjiang He, Yongxian Li, Luwen Pu, Huinan Xu, Chao Wang, and Shuangyong Song. 2024h. TeleChat: An open-source billingual large language model. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 10–20, Bangkok, Thailand. Association for Computational Linguistics.

Julia Watson, Sophia S. Lee, Barend Beekhuizen, and Suzanne Stevenson. 2025. Do language models practice what they preach? examining language ideologies about gendered language reform encoded in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1201–1223, Abu Dhabi, UAE. Association for Computational Linguistics.

Laura Weidinger, John F J Mellor, Bernat Guillén Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Diaz, A. Stevie Bergman, Mikel D. Rodriguez, Verena Rieser, and William Isaac. 2024. STAR: SocioTechnical approach to red teaming language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21516–21532, Miami, Florida, USA. Association for Computational Linguistics.

Xueru Wen, Jie Lou, Zichao Li, Yaojie Lu, XingYu XingYu, Yuqiu Ji, Guohai Xu, Hongyu Lin, Ben He, Xianpei Han, Le Sun, and Debing Zhang. 2025. Cheems: A practical guidance for building and evaluating Chinese reward models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15187–15211, Vienna, Austria. Association for Computational Linguistics.

Nevan Wichers, Carson Denison, and Ahmad Beirami. 2024. Gradient-based language model red teaming. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2862–2881, St. Julian's, Malta. Association for Computational Linguistics.

Erich H. Witte, Adrian Stanciu, and Klaus Boehnke. 2020. A new empirical approach to intercultural comparisons of value preferences based on schwartz's theory. *Frontiers in Psychology*, 11.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.

Di Wu, Jia-Chen Gu, Fan Yin, Nanyun Peng, and Kai-Wei Chang. 2024a. Synchronous faithfulness monitoring for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9390–9406, Miami, Florida, USA. Association for Computational Linguistics.

Jiayi Wu, Hengyi Cai, Lingyong Yan, Hao Sun, Xiang Li, Shuaiqiang Wang, Dawei Yin, and Ming Gao. 2025a. PA-RAG: RAG alignment via multi-perspective preference optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9091–9112, Albuquerque, New Mexico. Association for Computational Linguistics.

Shujin Wu, Yi R. Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025b. Aligning LLMs with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenya Wu and Weihong Deng. 2025. What counts underlying LLMs' moral dilemma judgments? In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 144–150, Vienna, Austria. Association for Computational Linguistics.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024b. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1332–1353, Miami, Florida, USA. Association for Computational Linguistics.

Yu Xia, Tong Yu, Zhankui He, Handong Zhao, Julian McAuley, and Shuai Li. 2024. Aligning as debiasing: Causality-aware alignment via reinforcement learning with interventional feedback. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4684–4695, Mexico City, Mexico. Association for Computational Linguistics.

Chaoyi Xiang, Chunhua Liu, Simon De Deyne, and Lea Frermann. 2025a. Comparing moral values in Western English-speaking societies and LLMs with word associations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3521–3536, Vienna, Austria. Association for Computational Linguistics.

Hao Xiang, Bowen Yu, Hongyu Lin, Keming Lu, Yaojie Lu, Xianpei Han, Ben He, Le Sun, Jingren Zhou, and Junyang Lin. 2025b. Self-steering optimization: Autonomous preference optimization for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9073–9085, Vienna, Austria. Association for Computational Linguistics.

Zeguan Xiao, Yan Yang, Guanhua Chen, and Yun Chen. 2024. Distract large language models for automatic jailbreak attack. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16230–16244, Miami, Florida, USA. Association for Computational Linguistics.

Guofu Xie, Xiao Zhang, Ting Yao, and Yunsheng Shi. 2025a. Bone soups: A seek-and-soup model merging approach for controllable multi-objective generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27237–27263, Vienna, Austria. Association for Computational Linguistics.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18128–18138, Miami,

Florida, USA. Association for Computational Linguistics.

Yiqing Xie, Wenxuan Zhou, Pradyot Prakash, Di Jin, Yuning Mao, Quintin Fettes, Arya Talebzadeh, Sinong Wang, Han Fang, Carolyn Rose, Daniel Fried, and Hejia Zhang. 2025b. Improving model factuality with fine-grained critique-based evaluator. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8140–8155, Vienna, Austria. Association for Computational Linguistics.

Bingbing Xu, Jing Yao, Xiaoyuan Yi, Aishan Maoliniyazi, Xing Xie, and Xiaofeng Meng. 2025a. Towards better value principles for large language model alignment: A systematic evaluation and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28991–29010, Vienna, Austria. Association for Computational Linguistics.

Qingting Xu, Kaisong Song, Chaoqun Liu, Yangyang Kang, Xiabing Zhou, Jun Lin, and Yu Hong. 2025b. COF: Adaptive chain of feedback for comparative opinion quintuple extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3236–3247, Abu Dhabi, UAE. Association for Computational Linguistics.

Rongwu Xu, Yishuo Cai, Zhenhong Zhou, Renjie Gu, Haiqin Weng, Liu Yan, Tianwei Zhang, Wei Xu, and Han Qiu. 2024a. Course-correction: Safety alignment using synthetic preferences. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1622–1649, Miami, Florida, US. Association for Computational Linguistics.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024b. Exploring multilingual concepts of human values in large language models: Is value alignment consistent, transferable and controllable across languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1771–1793, Miami, Florida, USA. Association for Computational Linguistics.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025c. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6859–6877, Albuquerque, New Mexico. Association for Computational Linguistics.

Shiyue Xu, Fu Zhang, Jingwei Cheng, and Linfeng Zhou. 2025d. MWPO: Enhancing LLMs performance through multi-weight preference strength and length optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20566–20581, Vienna, Austria. Association for Computational Linguistics.

Weiwen Xu, Deng Cai, Zhisong Zhang, Wai Lam, and Shuming Shi. 2024c. Reasons to reject? aligning language models with judgments. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12288–12304, Bangkok, Thailand. Association for Computational Linguistics.

Wenda Xu, Jiachen Li, William Yang Wang, and Lei Li. 2024d. BPO: Staying close to the behavior LLM creates better online LLM alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11125–11139, Miami, Florida, USA. Association for Computational Linguistics.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024e. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5587–5605, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. 2024f. A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7432–7449, Bangkok, Thailand. Association for Computational Linguistics.

Boyang Xue, Weichao Wang, Hongru Wang, Fei Mi, Rui Wang, Yasheng Wang, Lifeng Shang,

Xin Jiang, Qun Liu, and Kam-Fai Wong. 2023. Improving factual consistency for knowledge-grounded dialogue systems via knowledge enhancement and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7829–7844, Singapore. Association for Computational Linguistics.

Hongyu Yang, Jiahui Hou, Liyang He, and Rui Li. 2025a. Multi-perspective preference alignment of LLMs for programming-community question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1667–1682, Abu Dhabi, UAE. Association for Computational Linguistics.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Sen Yang, Nigel Collier, Dong Yu, and Deqing Yang. 2025b. LoGU: Long-form generation with uncertainty expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18947–18968, Vienna, Austria. Association for Computational Linguistics.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025c. Implicit cross-lingual rewarding for efficient multilingual preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21125–21147, Vienna, Austria. Association for Computational Linguistics.

Yan Yang, Zeguan Xiao, Xin Lu, Hongru Wang, Xuetao Wei, Hailiang Huang, Guanhua Chen, and Yun Chen. 2025d. SeqAR: Jailbreak LLMs with sequential auto-generated characters. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 912–931, Albuquerque, New Mexico. Association for Computational Linguistics.

Jing Yao, Xiaoyuan Yi, Shitong Duan, Jindong Wang, Yuzhuo Bai, Muhua Huang, Yang Ou, Scarlett Li, Peng Zhang, Tun Lu, Zhicheng Dou, Maosong Sun, James Evans, and Xing Xie. 2025. Value compass benchmarks: A comprehensive, generative and self-evolving platform for LLMs' value evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 666–678, Vienna, Austria. Association for Computational Linguistics.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024a. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8762–8785, Mexico City, Mexico. Association for Computational Linguistics.

Wenjin Yao, Yidong Wang, Zhuohao Yu, Rui Xie, Shikun Zhang, and Wei Ye. 2024b. PURE: Aligning LLM via pluggable query reformulation for enhanced helpfulness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8721–8744, Miami, Florida, USA. Association for Computational Linguistics.

Hai Ye and Hwee Tou Ng. 2024. Preference-guided reflective sampling for aligning language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21646–21668, Miami, Florida, USA. Association for Computational Linguistics.

Haoran Ye, TianZe Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. 2025. Generative psycho-lexical approach for constructing value systems in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11968–11991, Vienna, Austria. Association for Computational Linguistics.

Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2024. On the vulnerability of safety alignment in open-access LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9236–9260, Bangkok, Thailand. Association for Computational Linguistics.

Erxin Yu, Jing Li, and Chunpu Xu. 2024. PopALM: Popularity-aligned language models for social media trendy response predic-

tion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12867–12878, Torino, Italia. ELRA and ICCL.

Tianshu Yu, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2025. Diverse AI feedback for large language model alignment. *Transactions of the Association for Computational Linguistics*, 13:392–407.

Chenchen Yuan, Zheyu Zhang, Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. 2025. Probabilistic aggregation and targeted embedding optimization for collective moral reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11151–11168, Vienna, Austria. Association for Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5628–5643, Mexico City, Mexico. Association for Computational Linguistics.

Dun Zeng, Yong Dai, Pengyu Cheng, Longyue Wang, Tianhao Hu, Wanshun Chen, Nan Du, and Zenglin Xu. 2024. On diversified preferences of large language model alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9194–9210, Miami, Florida, USA. Association for Computational Linguistics.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. 2024. Removing RLHF protections in GPT-4 via fine-tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 681–687, Mexico City, Mexico. Association for Computational Linguistics.

Damin Zhang, Yi Zhang, Geetanjali Bihani, and Julia Rayz. 2025a. Hire me or not? examining language model's behavior with occupation attributes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7891–7911, Abu Dhabi, UAE. Association for Computational Linguistics.

Hangfan Zhang, Zhimeng Guo, Huaisheng Zhu, Bochuan Cao, Lu Lin, Jinyuan Jia, Jinghui Chen, and Dinghao Wu. 2024a. Jailbreak open-sourced large language models via enforced decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5475–5493, Bangkok, Thailand. Association for Computational Linguistics.

Jianfei Zhang, Jun Bai, Bei Li, Yanmeng Wang, Rumei Li, Chenghua Lin, and Wenge Rong. 2025b. Disentangling preference representation and text generation for efficient individual preference alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4813–4839, Abu Dhabi, UAE. Association for Computational Linguistics.

Jinchuan Zhang, Yan Zhou, Yaxin Liu, Ziming Li, and Songlin Hu. 2024b. Holistic automated red teaming for large language models through top-down test case generation and multi-turn interaction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13711–13736, Miami, Florida, USA. Association for Computational Linguistics.

Jingyu Zhang, Marc Marone, Tianjian Li, Benjamin Van Durme, and Daniel Khashabi. 2025c. Verifiable by design: Aligning language models to quote from pre-training data. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3748–3768, Albuquerque, New Mexico. Association for Computational Linguistics.

Mozhi Zhang, Pengyu Wang, Chenkun Tan, Mianqiu Huang, Dong Zhang, Yaqian Zhou, and Xipeng Qiu. 2025d. MetaAlign: Align large language models with diverse preferences during inference time. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5827–5845, Albuquerque, New

Mexico. Association for Computational Linguistics.

Xiaotian Zhang, Ruizhe Chen, Yang Feng, and Zuozhu Liu. 2025e. Persona-judge: Personalized alignment of large language models via token-level self-judgment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5037–5049, Vienna, Austria. Association for Computational Linguistics.

Xinliang Frederick Zhang, Winston Wu, Nick Beauchamp, and Lu Wang. 2024c. MOKA: Moral knowledge augmentation for moral event extraction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4481–4502, Mexico City, Mexico. Association for Computational Linguistics.

Yuqi Zhang, Liang Ding, Lefei Zhang, and Dacheng Tao. 2025f. Intention analysis makes LLMs a good jailbreak defender. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2947–2968, Abu Dhabi, UAE. Association for Computational Linguistics.

Wei Zhao, Zhe Li, Yige Li, Ye Zhang, and Jun Sun. 2024. Defending large language models against jailbreak attacks via layer-specific editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5094–5109, Miami, Florida, USA. Association for Computational Linguistics.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. MPO: Multilingual safety alignment via reward gap optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23564–23587, Vienna, Austria. Association for Computational Linguistics.

Wenliang Zheng, Sarkar Snigdha Sarathi Das, Yusen Zhang, and Rui Zhang. 2025. GreaterPrompt: A unified, customizable, and high-performing open-source toolkit for prompt optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 405–415, Vienna, Austria. Association for Computational Linguistics.

Enyu Zhou, Rui Zheng, Zhiheng Xi, Songyang Gao, Xiaoran Fan, Zichu Fei, Jingting Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. RealBehavior: A framework for faithfully characterizing foundation models' human-like behavior mechanisms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10262–10274, Singapore. Association for Computational Linguistics.

Guanghao Zhou, Panjia Qiu, Cen Chen, Hongyu Li, Jason Chu, Xin Zhang, and Jun Zhou. 2025a. LSSF: Safety alignment for large language models through low-rank safety subspace fusion. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30621–30638, Vienna, Austria. Association for Computational Linguistics.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. 2024a. WPO: Enhancing RLHF with weighted preference optimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8328–8340, Miami, Florida, USA. Association for Computational Linguistics.

Yue Zhou and Barbara Di Eugenio. 2025. Veracity bias and beyond: Uncovering LLMs' hidden beliefs in problem-solving reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21298–21310, Vienna, Austria. Association for Computational Linguistics.

Yukai Zhou, Jian Lou, Zhijie Huang, Zhan Qin, Sibei Yang, and Wenjie Wang. 2025b. Don't say no: Jailbreaking LLM by suppressing refusal. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25224–25249, Vienna, Austria. Association for Computational Linguistics.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao.

2024b. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10586–10613, Bangkok, Thailand. Association for Computational Linguistics.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024c. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10586–10613, Bangkok, Thailand. Association for Computational Linguistics.

Mingye Zhu, Yi Liu, Lei Zhang, Junbo Guo, and Zhendong Mao. 2024. LIRE: listwise reward enhancement for preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3377–3394, Bangkok, Thailand. Association for Computational Linguistics.

Elaine Zosa, Ville Komulainen, and Sampo Pyysalo. 2025. Got compute, but no data: Lessons from post-training a Finnish LLM. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 826–832, Tallinn, Estonia. University of Tartu Library.

Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024. A LLM-based ranking method for the evaluation of automatic counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9572–9585, Miami, Florida, USA. Association for Computational Linguistics.

A Theoretical Frameworks

We provide a comprehensive list of widely used frameworks for conceptualizing and/or operationalizing values, norms, and morals across levels of analysis in Table 4.

Level of analysis	**Frameworks**
Individual (micro)	Rokeach's Value Survey (RVS); Social Value Orientation (SVO); Kohlberg's Theory of Moral Development (TMD); Social Domain Theory (moral–conventional–personal); Rest's Four-Component Model; Moral identity; Moral disengagement
Individual ↔ Universal	Schwartz's Theory of Basic Human Values (TBHV); Moral Foundations Theory (MFT); Dual-process / social-intuitionist models of moral judgment (incl. Greene; Haidt)
Individual ↔ Norms (situational / interactional)	Theory of Planned Behavior / Reasoned Action (TPB/TRA: subjective norms, perceived control); Cialdini's Focus Theory of Normative Conduct (descriptive vs. injunctive norms, salience); Bicchieri's social norms framework (empirical + normative expectations; conditional preferences)
Cultural (macro)	Hofstede's Cultural Dimensions Theory (CDT); World Values Survey (WVS) / Inglehart–Welzel cultural dimensions; Cultural tightness–looseness
Cultural ↔ Universal	Inglehart's Post-Materialist Thesis / Human Development Sequence (HDS)
Institutional / collective action (meso)	Social norm enforcement, sanctioning, and informal social control (e.g., reputation, punishment); Ostrom's Institutional Analysis and Development (IAD) / rules-in-use
Individual ↔ Cultural (bridging typology)	De Groot & Steg (egoistic–altruistic–biospheric value orientations)
Values → Personal norms → behavior (bridging)	Norm Activation Model (NAM); Value–Belief–Norm (VBN) theory
Universal (normative)	Gert's Common Morality; Principlism in applied ethics/bioethics (autonomy, beneficence, nonmaleficence, justice)

Table 4: Selection of widely used frameworks for conceptualizing and/or operationalizing values, norms, and morals across levels of analysis. Here, *level of analysis* refers to the main social or conceptual scale a framework addresses, from individual preferences and moral judgments to cultural patterns and universal normative claims. Bridging frameworks connect more than one level.

B Methods

B.1 Retrieval and Automatic Filtering

The ACL Anthology Corpus filtering is detailed in Algorithm 1. This process resulted in 1,089 papers. The initial topic clustering for further automatic filtering is detailed in Algorithm 2. The resulting 20 clusters are displayed in Table 6. After manual screening, we excluded papers in clusters 1, 2, 5, 6, 9, 13, 14, 15, 16, and 17, as these clusters primarily focused on topics other than alignment. This process eliminated 373 entries and resulted in 790 papers that required manual annotation.

B.2 Manual Annotation

Annotators We employed three in-house annotators for this task. The annotators were from

keyword	# in titles	# in abstracts
align	811	3344
value	153	1301
moral	59	120
feedback	223	1005
reinforc	185	757
prefer	259	1254
sociotech	0	4

Table 5: Statistics of keyword occurrences in ACL Anthology papers published between 2022 and August 5, 2025.

Algorithm 1 ACL Anthology Filtering Pipeline

1: Download ACL Anthology entries from https://aclanthology.org/
2: Filter for entries published in or after 2022
Remaining entries: 37,260
3: Filter for entries containing an abstract
Remaining entries: 34,719
4: Filter for entries with English-language abstracts
Remaining entries: 34,135
5: Add keyword-match indicators for `align`, `value`, `moral`, `feedback`, `reinforc`, `prefer`, and `sociotech`
6: Retain entries with at least two keyword matches in the title or abstract
Remaining entries: 1,089

the international program of M.Sc. students in Computational Linguistics in their 3+ semesters, with excellent English competency and sufficient knowledge in the focus field. They were compensated at the state-level hourly salary of research assistants with bachelors degree.

To ensure strong inter-annotator agreement, the three annotators first labeled 100 instances. Agreement was assessed using Cohen's kappa in two rounds of 50 instances each. After each round, the annotators discussed disagreements and reached a consensus on the disputed items, during which they were given feedback on the annotations and provided with additional clarifications. Following the first round, agreement improved by an overall Δ of $+0.21$, with $+0.25$ on `alignment` and $+0.16$ on `social aspect`. Given this clear improvement, we proceeded with individual annotation of the remaining instances after the second-round discussion.

Codebook The annotators were given the items to annotate in individual CSV files and were given the codebook. They annotated the papers based on titles and abstracts on dimensions of `alignment` (TRUE/FALSE), `social aspect` (TRUE/FALSE), `modality` (Text, Image, etc.; multi-label possible), `language` (English, Indonesian, Multilingual, etc.; multi-label possible), `alignment target` (safety, bias, etc.; multi-label possible), and an optional field for notes (instructed to take notes if needed, especially if they found the item difficult to annotate).

ID	Cluster Name	#
-1	Mixed / residual topics	285
0	LLM safety, jailbreaks, and defenses	53
~~1~~	Medical and clinical LLM applications	22
~~2~~	Vision-language / multimodal alignment and evaluation	69
3	Moral reasoning, ethics, and values in LLMs	42
4	Retrieval, RAG, search, and ranking	42
~~5~~	Agents, environments, and sequential decision-making	27
~~6~~	Recommender systems and user preference modeling	27
7	Persona-aware and interactive dialogue systems	33
8	Persuasion, opinion, media, and social communication	18
~~9~~	Code generation and coding assistants	14
10	Cultural alignment and cross-cultural values	22
11	Social bias, stereotypes, and fairness	14
12	Linguistic structure and psycholinguistic analysis	12
~~13~~	Machine translation and multilingual generation	33
~~14~~	NLG evaluation, controllable generation, and metrics	24
~~15~~	Summarization and summary evaluation	22
~~16~~	Reasoning, search-augmented inference, and self-improvement	38
~~17~~	Education, tutoring, and feedback for learning	22
18	Instruction tuning, fine-tuning, and alignment data	58
19	Preference optimization, RLHF, DPO, and reward modeling	211

Table 6: Clusters obtained via BERTopic modeling with UMAP embeddings of filtered ACL Anthology abstracts. The papers excluded after this step are displayed as ~~strikeout~~.

Round	Annotator Pair	Alignment	Social Aspect
Round 1	Annotator 1 & 2	0.79	0.76
	Annotator 2 & 3	0.45	0.44
	Annotator 1 & 3	0.42	0.52
	Mean	**0.55**	**0.57**
Round 2	Annotator 1 & 2	0.96	0.92
	Annotator 2 & 3	0.66	0.59
	Annotator 1 & 3	0.79	0.67
	Mean	**0.80**	**0.73**

Table 7: Inter-annotator agreement across two rounds for the `alignment` and `social aspect` annotations, measured with Cohen's κ. The agreement Δ from Round 1$\rightarrow$Round 2 measured at 0.21 overall.

Annotation Outcome The annotators manually reviewed 790 paper titles and abstracts, and identified 497 alignment-focused papers, of which 291 addressed social aspects (Figure 3). Ten of these papers involved modalities beyond text only and were therefore excluded from our analysis. The final set of sociotechnical alignment papers thus comprised 281 papers.

Algorithm 2 Topic Modeling

function BUILD_TOPIC_MODEL(abstracts, random_state)
  Input: abstracts, and random state.
  Output: Topic assignments, probabilities, and trained topic model.
  $embedding_model \leftarrow$ SENTENCETRANSFORMER(model_name = all-MiniLM-L6-v2)
  $umap_model \leftarrow$ UMAP(min_dist = 0.5, metric = cosine, n_neighbors = 2, n_components = 2, low_memory = False, random_state = random_state)
  $topic_model \leftarrow$ BERTOPIC(umap_model = $umap_model$, embedding_model = $embedding_model$, verbose = True, nr_topics = 20, language = en, calculate_probabilities = True)
  $stop_words \leftarrow$ ENGLISHSTOPWORDS()
  for each abstract a in abstracts **do**
    $normalized_abstract \leftarrow$ REMOVESTOPWORDS(text = a, stop_words = $stop_words$)
    STORE(value=normalized_abstract, in=sociotech_topics)
  end for
  $(topics, probabilities) \leftarrow$ FITTRANSFORM(model = $topic_model$, documents = $normalized_abstracts$)
  return $topics, probabilities, topic_model$
end function

Algorithm 3 Automatic topical clustering

function TOPIC_MODELING(abstracts, n_topics, n_top_words)
  Input: abstracts, number of topics, and number of top words.
  Output: Top terms per topic and dominant topic assignments.
  $normalized_abstracts \leftarrow$ NORMALIZETEXT(texts = $abstracts$)
  $X \leftarrow$ TFIDF(texts = $normalized_abstracts$, stop_words = english, ngram_range=(1,2), min_df=2, max_df = 0.9, max_features = 4000)
  $(W, H) \leftarrow$ NMF(matrix = X, n_components = n_topics, init = nndsvda, random_state = 42, max_iter = 500)
  $topic_top_terms \leftarrow$ TOPTERMS(matrix = H, n_top_words = n_top_words)
  $dominant_topic \leftarrow$ ARGMAX(matrix = W, axis = 1)
  return $topic_top_terms, dominant_topic$
end function

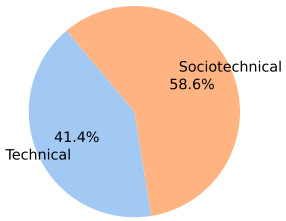

Figure 3: Distribution of purely "technical" vs. "socotechnical" alignment papers in the corpus: 206 technical alignment and 291 sociotechnical alignment, in total 497 alignment focused papers.

B.3 Automatic topical clusters

We detail the automatic topical clustering in Algorithm 3 (parameters are provided along with the algorithm). To obtain these clusters, we use scikit-learn's implementation of Non-Negative Matrix Factorization (NMF) together with TF–IDF vectorization from the same library. Table 3 provides an overview of the clusters.

C Literature References

We list all the literature in our corpus in Table 9.

Assigned Label	#	Consolidated Label
UNDERSPECIFIED	69	→ UNDERSPECIFIED (69)
SAFETY	62	→ SAFETY (62)
PERSONALIZATION	31	→ PERSONALIZATION (31)
CULTURAL	30	→ CULTURAL (33)
FACTUALITY	21	→ FACTUALITY&FAITHFULNESS (22)
MORAL	20	→ MORAL&ETHICAL (26)
BIAS	19	→ BIAS&FAIRNESS (19)
SOCIAL	18	→ SOCIAL&SOCIODEMOGRAPHIC (32)
VALUE	15	→ VALUE (15)
DEMOGRAPHICS	14	→ SOCIAL&SOCIODEMOGRAPHIC (32)
DIVERSITY	10	→ DIVERSITY (10)
MULTILINGUAL	6	→ LINGUISTICS&MULTILINGUAL (10)
ETHICAL	6	→ MORAL&ETHICAL (26)
POLITICAL	6	→ POLITICAL (6)
TOXICITY	6	→ TOXICITY (9)
LANGUAGE	4	→ LINGUISTICS&MULTILINGUAL (10)
OPINIONS	4	→ OPINIONS (4)
CULTURE	3	→ CULTURAL (33)
LEGAL	1	→ LEGAL (1)
HATE	1	→ TOXICITY (9)
HUMOR	1	→ HUMOR (1)
FAITHFULNESS	1	→ FACTUALITY&FAITHFULNESS (22)
SEXISM	1	→ TOXICITY (9)
LENGTH	1	→ LENGTH (1)
OFFENSIVENESS	1	→ TOXICITY (9)

Table 8: Annotated alignment target labels. The annotators were instructed to label papers without a clear alignment target as *UNDERSPECIFIED* and were allowed to assign multiple labels to individual papers.

ID	Clustered Literature
T0	Meister et al. (2025); Murthy et al. (2025); Galarnyk et al. (2025); Chen et al. (2025b); Spangher et al. (2024); Lerner et al. (2024); Hu et al. (2023); Moorjani et al. (2022); Huang and Xiong (2024); He et al. (2024); Liu et al. (2024a); Xu et al. (2024c); Hsu et al. (2024); Malik et al. (2024); Pandita et al. (2024); Ma et al. (2024); Zubiaga et al. (2024); Lee et al. (2024b); Vijjini et al. (2024); Wright et al. (2024); Taubenfeld et al. (2024); Feng et al. (2024); Potter et al. (2024); Nghiem et al. (2024); Fulay et al. (2024); Wang et al. (2024a); Dönmez and Falenska (2025); Asad et al. (2025); Watson et al. (2025); Zhang et al. (2025a); Gopalakrishna Pillai et al. (2025); Kirstein et al. (2025); Pugachev et al. (2025); Rennard et al. (2025); Park et al. (2025); Li et al. (2025b); Chen and Goldfarb-Tarrant (2025); Zhou and Di Eugenio (2025); Sun et al. (2025); Kim (2025); Kour et al. (2025); Ayele et al. (2024); Ceron et al. (2024); Kruspe (2024); Jones et al. (2024)
T1	Yu et al. (2025); Fu et al. (2025); Wang et al. (2025a); Kumar et al. (2025); Bu et al. (2025); Zhang et al. (2025d); Mire et al. (2025); Barnhart et al. (2025); Wang et al. (2024c); Liu et al. (2024d); Ryan et al. (2024); Khatun and Brown (2023); Dong et al. (2023); Kirk et al. (2023); Rosset et al. (2023); Zhu et al. (2024); Park et al. (2024); Hu et al. (2024b); Zhou et al. (2024c); Hong et al. (2024); Yao et al. (2024b); Zeng et al. (2024); Lu et al. (2024a); Guo et al. (2024b); Chen et al. (2024c); Cao et al. (2024b); Kwon et al. (2024); Xie et al. (2024); Lloret et al. (2024); Song et al. (2025a); Xu et al. (2025d); Wang et al. (2025c); Riahi Samani et al. (2025); Li et al. (2025a); Huang et al. (2025); Xie et al. (2025a); Hamilton (2024); Xia et al. (2024); Zhan et al. (2024)
T2	Roberts et al. (2025); Harðarson et al. (2025); Nath et al. (2025b); Zhang et al. (2025c); Wu et al. (2025a); Lee and Han (2025); Wadhwa et al. (2025); Ye and Ng (2024); Padhi et al. (2024); Huang et al. (2023); Lu et al. (2023); Naseem et al. (2024); Cheng et al. (2024); Nguyen et al. (2024a); Amini et al. (2024); Fan et al. (2024); Duan et al. (2024); Wu et al. (2024b); Zhou et al. (2024a); Xu et al. (2024d); Zhang et al. (2025e); Xiang et al. (2025b); Xu et al. (2025b); Zhang et al. (2025b); Takayama et al. (2025); Alnumay et al. (2025); Nath et al. (2025a); Wen et al. (2025); Guo et al. (2025a); Corrado et al. (2025); Wang et al. (2025f,d); Qiyuan et al. (2025); Thorne et al. (2024); Yue et al. (2024); Shi et al. (2024)
T3	Yang et al. (2025d); Biancotti et al. (2025); Jiang et al. (2025c); Hazra et al. (2024); Wichers et al. (2024); Wang et al. (2024e); Zhang et al. (2024a); Xu et al. (2024e); Cao et al. (2024a); Xu et al. (2024a,f); Yi et al. (2024); Ren et al. (2024); Chehbouni et al. (2024); Zhao et al. (2024); Liu et al. (2024b); Xiao et al. (2024); Kumarage et al. (2025); Ganon et al. (2025); Zhou et al. (2025b); Zhang et al. (2025f); Li et al. (2025d); Liu et al. (2025e); Lee et al. (2025); Zhao et al. (2025); Guo et al. (2025b); Zhou et al. (2025a); Ji et al. (2025a); Rebedea et al. (2025); Luo et al. (2024); Chen et al. (2024a); Cao et al. (2024d)
T4	Mittal et al. (2025); Cahyawijaya et al. (2025); Bang et al. (2023); Zhou et al. (2023); Qiu et al. (2022); Ammanabrolu et al. (2022); Liu et al. (2022); Al Ali and Libovický (2024); Lu et al. (2024b); Lee et al. (2024c); Liu et al. (2024c); Ju et al. (2025); Kang et al. (2025); Marco et al. (2025); Long et al. (2025); Dong et al. (2025); Liu et al. (2025d); Jiang et al. (2025a); Ye et al. (2025); Han et al. (2025); Shetty et al. (2025); Jin et al. (2025); Xu et al. (2025a); Choi et al. (2025); Yao et al. (2025); Huang et al. (2024b); Yao et al. (2024a); Hu et al. (2024a)
T5	Skenderi et al. (2025); Duan et al. (2025); Baskar et al. (2025); Mishra et al. (2024); Raju et al. (2024); Wang et al. (2024b); Yu et al. (2024); Li et al. (2024); Chen et al. (2024d); Wang et al. (2024d); Zhang et al. (2024b); Guan et al. (2025); Ji et al. (2025c); Yang et al. (2025a); Wu et al. (2025b); Castricato et al. (2025); Karine and Marlin (2025); Qin et al. (2025); Balepur et al. (2025); Hoyle et al. (2025); Wang et al. (2025b); Hao et al. (2025); Oh et al. (2025); Zheng et al. (2025); Lin et al. (2024)
T6	Liu et al. (2025b); Hakim et al. (2025); Wu and Deng (2025); Leteno et al. (2025); Marraffini et al. (2024); Paulissen and Wendt (2023); Rao et al. (2023); Ramezani and Xu (2023); Sun et al. (2023); Fraser et al. (2022); Guan et al. (2022); Liu et al. (2024f); Hobson et al. (2024); Abdulhai et al. (2024); Davani et al. (2024); Guo et al. (2024a); Yuan et al. (2025); Xiang et al. (2025a); Greco et al. (2025); Falk and Lapesa (2025); Alvarez Nogales and Araque (2024); Zhang et al. (2024c)
T7	Kravchenko et al. (2025); Chen et al. (2025a); Zosa et al. (2025); Sharma et al. (2025a); Corral et al. (2025); Song et al. (2025b); Jinnai (2024); Ahmad et al. (2024); Nguyen et al. (2024b); Lai et al. (2023); España-Bonet and Barrón-Cedeño (2022); Agarwal et al. (2024); Shen et al. (2024); Xu et al. (2024b); Aakanksha et al. (2024); Dang et al. (2024); Yang et al. (2025c); Liu et al. (2025c); Kim and Kim (2025); Wang et al. (2024h); Chia et al. (2024)
T8	Baek et al. (2025); Wang et al. (2025e); Xue et al. (2023); Liang et al. (2024); Chen et al. (2024b); Das and Srihari (2024); Tao et al. (2024); Rozner et al. (2024); Lee et al. (2024a); Wu et al. (2024a); Liu et al. (2024e); Ko et al. (2024); Bi et al. (2025); Xie et al. (2025b); Yang et al. (2025b); Li et al. (2025c); Du et al. (2024)
T9	Xu et al. (2025c); Banerjee et al. (2025); Magdy et al. (2025); Nacar et al. (2025); Wang et al. (2024g); AlKhamissi et al. (2024); Arora et al. (2023); Cao et al. (2024c); Acquaye et al. (2024); Hosseinbeigi et al. (2025); Masoud et al. (2025); Pandey et al. (2025); Keleg (2025); Liu et al. (2025a); Feng et al. (2025); Huang et al. (2024a)

Table 9: The sociotechnical alignment corpus. The references listed here correspond to the literature in Table 3, which is summarized in Section 4.2.